Model for Monitoring and Assessment of a Public Health Care Network

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia, May 2011

Supervisor: Prof. Dr. Marko Bohanec, Jožef Stefan Institute, Ljubljana, Slovenia

Evaluation Board:
Prof. Dr. Nada Lavrač, Jožef Stefan Institute, Ljubljana, Slovenia
Prof. Dr. Sašo Džeroski, Jožef Stefan Institute, Ljubljana, Slovenia
Prof. Dr. Joost N. Kok, Leiden University, The Netherlands

Aleksander Pur

MODEL FOR MONITORING AND ASSESSMENT OF A PUBLIC HEALTH CARE NETWORK

Doctoral Dissertation

MODEL ZA SPREMLJANJE IN VREDNOTENJE JAVNE ZDRAVSTVENE MREŽE

Doktorska disertacija

Supervisor: Prof. Dr. Marko Bohanec

June 2011

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL
Ljubljana, Slovenia

Contents

Abstract

It is hard to imagine the management of any system without monitoring it as a process of continuously gathering data and performing real-time analyses. In general, monitoring can improve the estimation of the current state, optimization of the business processes, identification of the critical elements and new opportunities, and prediction of the future state and planning. This dissertation is focused on the development of models for the monitoring of organizational systems such as the Primary Health Care Network (PHCN). In order to improve the monitoring systems, this work proposes the use of advanced data analysis techniques such as assessment models, Data Mining (DM) techniques, and advanced data visualizations. Considering that these techniques can make monitoring models very complicated, we suggest a new methodology for developing monitoring models based on Hierarchical Assessment and Monitoring Models (HAMM), the Generic Templates (GT), and advanced data analysis techniques. The contributions of this dissertation also include suggestions for using advanced data analysis techniques in monitoring systems, the new model for monitoring of PHCNs, and analyses of the Slovenian PHCN based on the proposed model in the projects of MediMap (2004), MediNet (2005), MediNet+ (2006), and MediNet++ (2008).

Povzetek

Težko si zamislimo uspešno upravljanje kateregakoli sistema brez stalnega spremljanja in ocenjevanja njegovih procesov, lastnosti in zakonitosti. V disertaciji smo se osredotočili na tehnike in modele za spremljanje in ocenjevanje procesov in lastnosti organizacijskih sistemov. Namen disertacije je izboljšanje teh modelov z uporabo naprednejših tehnik analize podatkov, ki smo jih razdelili v modele za ocenjevanje, metode rudarjenja podatkov in naprednejše tehnike vizualizacije podatkov. Eden od glavnih ciljev disertacije je izboljšan pristop k izdelavi tovrstnih modelov z uporabo naprednejših tehnik za analizo podatkov. Predlagan pristop temelji na hierarhičnih modelih, splošnih predlogah za izgradnjo modelov in uporabi naprednejših tehnik za analizo podatkov. Največjo uporabnost naprednejših tehnik smo dosegli pri ocenjevanju stanja in kritičnih pojavov v sistemu. Organizacijski sistem, na katerega smo se pretežno omejili v disertaciji, je mreža javnih zdravstvenih služb (MJZS). Glavni prispevki disertacije so nov koncept za izdelavo tovrstnih modelov, model za spremljanje mreže javnih zdravstvenih služb in konkretne analize v okviru projektov MediMap (2004), MediNet (2005), MediNet+ (2006), in MediNet++ (2008).

Uvod

Širše področje disertacije so tehnike in modeli za spremljanje in ocenjevanje procesov in lastnosti organizacijskih sistemov. Kljub temu, da običajno nismo preveč navdušeni nad tem, da bi nekdo stalno spremljal in ocenjeval naše aktivnosti, si težko zamislimo uspešno upravljanje kateregakoli sistema brez stalnega spremljanja in ocenjevanja njegovih procesov, lastnosti in zakonitosti. Praviloma to velja tako za mehanske in biološke kot tudi za organizacijske sisteme. V disertaciji je proces spremljanja in ocenjevanja sestavljen iz periodičnega zbiranja podatkov o sistemu in njihovih analiz z namenom zagotavljanja koristnih informacij tistim, ki so odgovorni za njegovo delovanje. Pridobljene informacije lahko pomagajo upravitelju sistema za oceno trenutnega stanja sistema, izboljšanje in optimiziranje njegovih procesov, prepoznavanje kritičnih elementov in tveganj v sistemu, boljše predvidevanje obnašanja sistema v prihodnosti itd.

Namen disertacije je izboljšanje teh modelov z uporabo naprednejših tehnik za analizo podatkov (poglavje 3), s katerimi odkrivamo zakonitosti v podatkih in so značilne za sisteme za podporo odločanju. V disertaciji smo naprednejše tehnike razdelili v modele za ocenjevanje, metode rudarjenja v podatkih in naprednejše tehnike vizualizacij podatkov. Njihovo največjo uporabnost smo dosegli pri ocenjevanju lastnosti in identifikaciji pomembnih pojavov v opazovanih sistemih (poglavje 5). Ker lahko uporaba tovrstnih tehnik poveča kompleksnost modelov za spremljanje organizacijskih sistemov, smo izdelati nov pristop za njihovo izgradnjo, ki temelji na hierarhičnih modelih (razdelek 4.1) splošnih predlogah za izgradnjo modelov (razdelek 4.2) in uporabi naprednejših tehnik za analize podatkov (poglavje 3).

Organizacijski sistem, na katerega smo se pretežno omejili v disertaciji, je mreža javnih zdravstvenih služb (MJZS). Javne službe so praviloma namenjene zadovoljevanju določenih skupnih potreb prebivalcev. Nekatere izmed njih, kot na primer policija, šolstvo in zdravstvo, so z namenom, da bi se bolj približale prebivalcem, svoje službe organizirale v organizacijske sisteme, ki jih imenujemo tudi mreže. Ena izmed njih je tudi MJZS, ki je namenjena pokrivanju potreb po zdravstvenem varstvu prebivalcev Slovenije in vseh drugih, ki iščejo zdravstveno pomoč. Za to mrežo je pomembna pravilna razporeditev človeških in materialnih virov javnih zdravstvenih zavodov ter koncesionarjev tako, da državljanom zagotavlja enakomerno in pravično dostopnost do zdravstvene oskrbe, ki je financirana iz javnih sredstev.

MJZS je glede na izvajalce razdeljena na primarno, sekundarno in terciarno raven. Primarno raven sestavljajo zdravstvene službe splošne medicine, pediatrije (zdravstveno varstvo otrok in mladine) in ginekologije (zdravstveno varstvo žensk). Sekundarna raven pa je sestavljena iz različnih specialnosti: interne medicine, splošne kirurgije, pediatrije, ginekologije, oftalmologije, nevrologije itd. Predstavnik terciarne ravni je Univerzitetni klinični center Ljubljana. Načrtovanje in spremljanje posameznih ravni mreže poteka ločeno. Še zlasti za primarno raven velja, da poteka načrtovanje in spremljanje MJZS pogosto brez enotnih meril in kriterijev ter je prepuščeno lokalnim oblastem. Takšno načrtovanje lahko pripelje do neenakosti pri razpoložljivost zdravstvenih storitev za prebivalce in neučinkovite izrabe zmogljivosti izvajalcev. Zaradi tega se Ministrstvo za zdravje RS zavzema za določitev enotnega modela za spremljanje in načrtovanje MJZS.

Glavni cilji disertacije so:

- izboljšati pristop k izdelavi modelov za spremljanje in ocenjevanje procesov v organizacijskih sistemih z uporabo naprednejših tehnik analize podatkov;
- uporaba predlaganega pristopa za izdelavo modelov za spremljanje konkretnih organizacijskih sistemov;
- izdelava modela za spremljanje MJZS, z uporabo naprednejših tehnik analize podatkov;
- izdelava konkretnih analiz MJZS na osnovi izdelanega modela.

Glavni prispevki disertacije so:

- predlog novega pristopa k razvoju modelov za spremljanje in ocenjevanje procesov in lastnosti organizacijskih sistemov (poglavje 4),
- nov koncept hierarhičnih modelov za ocenjevanje in spremljanje sistemov (razdelek 4.1);
- uporaba splošnih predlog za izgradnjo tovrstnih modelov (razdelek 4.2),

- uporaba naprednejših tehnik analiz podatkov v predlaganih modelih predvsem za ocenjevanje procesov in lastnosti organizacijskih sistemov (poglavje 3),
- nov model za spremljanje MJZS (poglavje 5),
- predlog novih meril in kriterijev za spremljanje MJZS (poglavje 5) in
- konkretne analize MJZS na osnovi predlaganega modela v okviru projektov MediMap (2004), MediNet (2005), MediNet+ (2006) in MediNet++ (2008).

V disertaciji smo predlagali nov pristop k izdelavi modelov za spremljanje raznovrstnih sistemov. Ta temelji na hierarhično povezanih modulih, ki prikazujejo različne aspekte MJZS. Vsak modul je sestavljen iz najmanj enega procesa, ki zagotavlja informacije, povezane z določenim aspektom mreže (razdelek 4.1).

Ugotovili smo, da so nekateri osnovni koncepti tovrstnih modelov uporabni tudi v podobnih modelih na drugih področjih. Te osnovne koncepte smo prav tako prikazali v obliki hierarhično povezanih modulov (razdelek 4.2).

Predlagani model za spremljanje MJZS zagotavlja koristne informacije, ki so povezane z načrtovanjem virov v MJZS predvsem z vidika uporabnika zdravstvenih storitev. Zaradi tega smo v model vključili raznovrstne metode analiz in vizualizacij podatkov (poglavje 3). Nekaj teh metod je bilo prvič uporabljenih prav v modelu za spremljanju MJZS v Sloveniji. Takšnega modela za spremljanje in načrtovanje MJZS, ki na predlagani način združuje tako široko množico metod, po naših informacijah ni niti v svetovnem merilu. V okviru disertacije smo prav tako razvili nova in izboljšali stara merila za spremljanje MJZS, ki prav tako pomenijo novost tudi v svetu (poglavje 5).

Bistveni prispevki disertacije so tudi konkretne analize podatkov MJZS v okviru projektov MediMap (2004), MediNet (2005), MediNet+ (2006) in MediNet++ (2008). Rezultati analiz so neposredno uporabni za Ministrstvo za zdravje Republike Slovenije kot načrtovalca zdravstvene mreže, za Inštitut za varovanje zdravja v smislu spremljanja zdravstvenega stanja prebivalcev ter za Zavod za zdravstveno zavarovanje Slovenije v smislu spremljanja in predvidevanja uporabe zdravstvenih zmogljivosti in financ.

Predvidevamo, da bodo vse predlagane metode in pristopi prispevali k razvoju širšega znanstvenega področja razvoja sistemov za spremljanje in ocenjevanje raznovrstnih sistemov. Prav tako lahko rezultati disertacije pomagajo pri načrtovanju podobnih modelov tako za druge javne servise kot sta šolstvo in policija, kot tudi za ostale sisteme kot je na primer sistem za vzdrževanje osnovnih sredstev.

Abbreviations

BI	=	Business Intelligence
DSS	=	Decision Support System
GT	=	Generic Template
HAMM	=	Hierarchical Assessment and Monitoring Models
HCN	=	Health Care Network
KDD	=	Knowledge Discovery in Databases
MM	=	Monitoring Model
MoH	=	Ministry of Health
MP	=	Monitoring Process
OLAP	=	On Line Analytical Process
PHCN	=	Primary Health Care Network

1 Introduction

If you cannot measure it, you cannot control it.
If you cannot control it, you cannot manage it.
If you cannot manage it, you cannot improve it.

H. J. Harrington

1.1 Motivation

Even though people do not want to be measured or monitored, the monitoring of any system is important for its management, and consequently for its improvement. In general, the monitored system can be any single organism, any organization or society, any electro-mechanical or informational artifact, human or animal body, living environment, etc. The term *monitoring* denotes the process of continuously gathering and performing real-time analyses of data about the monitored system in order to get information that supports its management. Thus, this information can help users to estimate the current state, to optimize the system's processes, to recognize the critical elements of the system and new opportunities, and to plan and improve prediction of the future state. For example, the turbulent business environment today requests many companies and public organizations to change their processes and organization structures. In order to discover how successful the organizational changes are as well as to identify the weakest links, assess risks, and predict future states, system's management requires a comprehensive monitoring system. The term *monitoring system* denotes the information system aimed at monitoring of some other system. In the dissertation, the basic concept and important details of a monitoring system are represented by *monitoring models*.

This dissertation is focused on the development of models for monitoring of the Primary Health Care Network (PHCN). However, the methodology, concepts, and techniques can be used for the monitoring of many other systems. We assume that all efficient and effective monitoring systems have to meet the following requirements:

- On the one hand, the system has to measure various performances about the monitored system. On the other hand, the system has to be able to assess the meaning of the measured values for the past, present, and future (Section 2.3).
- The monitoring system has to provide all relevant expected and unexpected information about the monitored system. In general, the expected information is provided by predefined indicators. At the same time, additional and sometimes unexpected information can be provided by Data Mining (DM) techniques, such as association rules discovery, clustering, and prediction (Chapter 3).
- The system has to induce relevant information from a large amount of data in real-time (as soon as possible) in order to prevent undesired consequences.
- The information has to be provided to users in an automated or semi-automated manner, and pushed rather than pulled.
- The information has to be presented in a clear, comprehensible, and unambiguous way, because the users are not always highly skilled in statistical and analytical techniques.

Considering these requirements, the monitoring systems have to include data analysis techniques providing basic metrics as well as advanced techniques for data analyses and visualization of results. The problem is that these monitoring systems can become huge and complicated (Wang, et al., 2009). For instance, the systems such as a Health Care Network (HCN) or those of a national police force usually generate daily large amount of data in various forms. The analyses of these data using various technologies reveal many aspects of the systems, such as human resources, allocation of health care resources, availability of the Health Care Providers

(HCP), the state of health of the population, and quality of the health care services. These aspects are related to each other, and each aspect can be revealed by many indicators. Moreover, the monitoring systems have to be able to provide the basic aspects as well a large number of important details. Thus, the development and management of these systems are inherently difficult.

In order to improve the functionality and comprehensibility of the monitoring models as well as the implementation of various data analysis methods, we propose a methodology for developing monitoring models (MM) that is based on three important concepts (Chapter 4): (1) the Hierarchical Assessment and Monitoring Model (HAMM), (2) the Generic Template (GT), and (3) advanced data analysis techniques.

The HAMM describes the basic concept and all details of the model. It is composed of hierarchically connected modules aimed at certain aspects of the monitored system. This methodology is not limited to any particular data analysis and data visualization techniques. Thus, the models are aimed at simple monitoring as well as assessments and predictions.

Some concepts described by the HAMM are general. These concepts can be used for monitoring of similar systems in various fields. We call these concepts the Generic Template (GT). These concepts can simplify the development of monitoring models and allow the reuse of similar models.

In the dissertation, the data analyses techniques for monitoring systems are divided into simple indicators and advanced data analysis techniques. On the one hand, the monitoring system continuously measures performances and events by various indicators. On the other hand, it can assess what the measured values mean for the current and future states of a monitored system. These assessments are usually provided by advanced data analysis techniques such as assessment models, Data Mining (DM) techniques, and advanced data visualizations.

1.2 Hypothesis and Purpose

The hypothesis of the dissertation is that advanced data analysis techniques can improve the monitoring and assessment of various systems. In the dissertation, the advanced data analysis techniques (Chapter 3) include assessment models, Data Mining (DM) techniques, and advanced data visualization techniques.

The monitoring systems that are composed of various simple indicators as well as the advanced data analysis techniques can be huge and complicated, and thus not easy to design. Moreover, the implementation of the advanced data analysis techniques requires a wide range of knowledge and skills.

The purpose of the dissertation is to improve the methodology to design of the MM, which is composed of various data analysis techniques. The dissertation is particularly focused on the model for monitoring of the Slovenian PHCN.

The main goals of the dissertation are:

- To improve the methodology for developing monitoring models including advanced data analysis techniques.
- To use the proposed methodology and advanced techniques in various monitoring models.
- To develop a monitoring and assessment model aimed at the PHCN.
- To analyze the Slovenian PHCN in accordance with the new PHCN monitoring and assessment models.

1.3 Scientific Contributions

The general contributions of this dissertation consist of the following:

- A proposal for a new methodology for developing monitoring models including monitoring and assessment processes (Chapter 4).
- HAMM – a hierarchical approach to the description of basic concepts and details of the monitoring models (Section 4.1).
- Generic Template – the reusable concepts of the monitoring models (Section 4.2).

- Implementation of the advanced data analysis techniques (Chapter 7) in MM that support assessment processes.

The contributions of this dissertation to the monitoring of the PHCN consist of the following:

- New indicators for monitoring of the PHCN (Chapter 5).
- A new PHCN monitoring and assessment model aimed at resource allocation (Chapter 5).
- Monitoring and assessment models aimed at the health care system based on the proposed methodology (Chapter 6).
- Application of proposed data analysis concepts and techniques for the analyses of real data performed within the projects described below.

The research, discoveries, and conclusions in the dissertation are mainly based on the work that was carried out in the following projects**:**

- MediMap: Knowledge Management in Medicine and Health Care (Analiza podatkov za upravljanje znanja na področju zdravstva); project funded by Health Care Institute Celje (2003-2004).
- MediNet: Analysis of Factors in Setting up a Network of Health Care Personnel (Analiza dejavnikov za postavitev mreže zdravstvenih delavcev); project funded by the Ministry of Health of the Republic of Slovenia (2005).
- MediNet+: Development of a Primary Health Care Network in Slovenia (Izdelava modela mreže zdravstvenih delavcev primarne ravni Slovenije); project funded by the Ministry of Health of the Republic of Slovenia (2006).
- MediNet++: A Web Page for Presenting the Primary Health Care Network in Slovenia (Internetna predstavitev rezultatov analiz za spremljanje mreže zdravstvenih delavcev primarne ravni v Sloveniji), project funded by the Ministry of Health of the Republic of Slovenia (2008).

The discoveries and conclusions related to the dissertation were published in the following publications**:**

Original Scientific Article

PUR, Aleksander, BOHANEC, Marko, CESTNIK, Bojan, LAVRAČ, Nada, DEBELJAK, Marko, KOPAČ, Tadeja. Data Mining for Decision Support: An Application in Public Health Care. *Lect. Notes Comput. Sci.*, 2005, vol. 3533, str. 459–469. [COBISS.SI-ID 19103271]

LAVRAČ, Nada, BOHANEC, Marko, PUR, Aleksander, CESTNIK, Bojan, JERMOL, Mitja, URBANČIČ, Tanja, DEBELJAK, Marko, KAVŠEK, Branko, KOPAČ, Tadeja. Resource Modeling and Analysis of Regional Public Health Care Data by Means of Knowledge Technologies. *Lect. Notes Comput. Sci.*, 2005, vol. 3581, str. 414–418. [COBISS.SI-ID 19201831]

LAVRAČ, Nada, BOHANEC, Marko, PUR, Aleksander, CESTNIK, Bojan, DEBELJAK, Marko, KOBLER, Andrej. Data Mining and Visualization for Decision Support and Modeling of Public Health Care Resources. *Journal of Biomedical Informatics*, 2007, vol. 40, no. 4, str. 438–447. [COBISS.SI-ID 20957479]

PUR, Aleksander, BOHANEC, Marko, LAVRAČ, Nada, CESTNIK, Bojan, DEBELJAK, Marko, GRADIŠEK, Anton. Monitoring Human Resources of a Public Health Care System Through Intelligent Data Analysis and Visualization. *Lect. Notes Comput. Sci.*, 2007, vol. 4594, str. 175–179. [COBISS.SI-ID 20900391] tipologija 1.08 -> 1.01

PUR, Aleksander, BOHANEC, Marko, LAVRAČ, Nada, CESTNIK, Bojan. Primary Health-care Network Monitoring: A Hierarchical Resource Allocation Modeling Approach. *Int. J. Health Plann. and Manage.*, 2010, vol. 25, no. 2, str. 119–135. [COBISS.SI-ID 23721255]

Professional Article

PUR, Aleksander, PRIBAKOVIĆ BRINOVEC, Radivoje, LAVRAČ, Nada, DEBELJAK, Marko, BOHANEC, Marko, CESTNIK, Bojan, URBANČIČ, Tanja, ALBREHT, Tit, KOPAČ, Tadeja, KOBLER, Andrej, KLEMENC, Jernej, LUKŠIČ, Primož. Sodobne metode analiziranja in vrednotenja primarne in sekundarne ravni zdravstvenega varstva. Bilt.-ekon. organ. inform. zdrav., 2006, letn. 22, št. 1, str. 4–17. [COBISS.SI-ID 19781415]

Published Scientific Conference Contribution

PUR, Aleksander, BOHANEC, Marko, LAVRAČ, Nada, CESTNIK, Bojan. Data Presentation Methods for Monitoring a Public Health Care System. V: JARM, Tomaž (ur.), KRAMAR, Peter (ur.), ŽUPANIČ, Anže (ur.). 11th Mediterranean Conference on Medical and Biological Engineering and Computing 2007, 26–30 June, 2007, Ljubljana, Slovenia, (IFMBE proceedings, vol. 16). New York: Springer: International Federation for Medical and Biological Engineering, 2007, str. 708–711. [COBISS.SI-ID 20942631]

Published Professional Conference Contribution Abstract

PUR, Aleksander. Presentation of the Concept of the Border Data Analyses System, Expert Meeting on Information Management, DCAF, Budva, May 2009.

PUR, Aleksander, PRIBAKOVIĆ BRINOVEC, Radivoje, LAVRAČ, Nada, BOHANEC, Marko, CESTNIK, Bojan, URBANČIČ, Tanja, ALBREHT, Tit, KOPAČ, Tadeja, LUKŠIČ, Primož. Sodobne metode analiziranja in vrednotenja primarne in sekundarne ravni zdravstvenega varstva = New methods in analyzing and evaluating primary and secondary health care. *Bilt.-ekon. organ. inform. zdrav.*, 2005, letn. 21, izredna št., str. 30–31. [COBISS.SI-ID 1462245]

Short Scientific Article

PUR, Aleksander, BOHANEC, Marko. Police Data Analyses with Data Mining Methods. *Varstvoslovje*, apr. 2003, letn. 5, št. 1, str. 16–21, graf. prikazi. [COBISS.SI-ID 670186]

1.4 Thesis Structure

This thesis is structured as follows. The motivation and hypothesis are presented in Chapter 1. Chapter 2 describes the background and main terminology. The advanced data analysis techniques are described Chapter 3. The description of the methodology for developing the HAMM is presented in Chapter 4. Chapter 5 presents proposed PHCN monitoring and assessment model. The main part of this thesis consists of five original research papers on monitoring models for the health care system co-authored by the author of this dissertation. They were published in internationally recognized journals and presented in Chapter 6. Chapter 7 shows some recommendations for implementation of the advanced data analysis techniques in MM7 . Chapter 8 discusses the results and concludes the thesis.

2 Background and Related Work

This chapter provides general information about monitoring processes and monitoring systems as well as monitoring of health care networks. In particular, it is focused on the assessment processes in monitoring systems supported by the advanced data analysis techniques. Theses processes try to assess and reveal the meaning of the measured values, i.e., indicators.

2.1 Monitoring Systems

Monitoring systems are information systems aimed at monitoring of various other systems, such as organization or society, electro-mechanical or informational artifact, human or animal body, and living environment. Basic elements of monitoring systems are monitoring processes. Each monitoring process reveals some aspect of the monitored system.

In the business world, Business Intelligence (BI) is widely used as a framework for the monitoring of organizations (Turban, et al., 2010). In general, the BI systems are Decision Support Systems (DSS) aimed at the support of business decision-making processes. Traditional BI systems include common elements such as reporting, querying, Key Performance Indicators, Online Analytical Processing (OLAP), dashboards, scorecards, and statistics. Nowadays the vendors are focused on the technical ability of their systems such as integration with other technologies, the way they compress data, and the speed with which their hash algorithms slice through data or their new rich graphics. On the one hand, this approach is quite appropriate when the analysis is based on clear indicators and structured data. On the other hand, considering data sources such as a large amount of unstructured data, social networks, WEB 2.0, geotagging data as well as data analysis techniques such as the data mining, text mining, real-time analyses, and advanced visualization techniques, some useful information can stay hidden when the organizational systems are monitored by the traditional BI systems. Thus, the new generation of Business Intelligence systems (BI 2.0) (Nelson, 2010) provide new functionalities such as:

- proactive alerts and notifications,
- improved visualizations,
- text mining,
- predictive analytics,
- BI as a service (SOA),
- in-memory analytics,
- open Source BI, and
- event-driven real-time access to information provided by Business Activity Monitoring (BAM) and Complex Event Processing (CEP).

In the realm of environment protection, the monitoring systems are focused on such aspects as the weather conditions, biotic diversity, and pollution. Usually, the prediction is one of the most important functionalities of these systems. For example, there is a system aimed at monitoring and predicting biological, ecological, and physical processes along the Gulf of Mexico coastal lines (Ladner, et al., 2009). There, the real-time monitoring and prediction system is based on the fusion of surface bio-optical properties of multiple satellites with physical models and observations. By using various data analysis methods, the system provides coastal zone managers with the capability to assess, predict, and track Harmful Algal Blooms, hypoxia, and sediment discharge. In general, such systems are characterized by various data gathering methods such as environment sensors, satellite data, and demographic data, as well as advanced data analysis methods that assess and predict critical situations.

2.2 Monitoring Processes

Systems monitoring can be decomposed into the following processes:

1. *Data gathering*: The gathering of data about a monitored system is based on various techniques such as surveys, automatic conversion of physical values into electronic data by sensors and satellites, speech and speaker recognitions, and automatic image analysis as well as collecting by standard procedures in organizations.

2. *Data preparation*: Data preparation is a process in which data are prepared for analyses by techniques such as data extraction, cleaning, joining, and transforming (Rahm and Hai, 2000). This process has to be done automatically as much as possible.

3. *Measuring of performances and events*: This process is aimed at the measuring of performances and events related to a monitored system. Inputs to this process are provided through various indicators. In this paper, the term indicator refers to models that measure size, quantity, performance, etc. Usually, this process is limited to simple and clear data operations calculated by Structured Query Language (SQL) formulas.

4. *Assessment*: This process assesses what the measured values mean for the current and future states of the monitored system. Assessment is not limited to any particular data analysis method, however, it is usually based on advanced data analysis techniques such as assessment models, data mining, and advanced data visualization methods.

5. *Dissemination of results*: The fifth process is focused on providing useful information to those who need it. The user can access information in pull or push mode. In pull mode, the information is provided on demand. In push mode, the information is automatically sent to users when they need it.

6. *Interpretation of results*: The interpretation is a phase in which the user interprets the information provided by the monitoring system in order to make appropriate decisions. The provided information has to be comprehensible, because the users are not always highly skilled in statistics and data analysis.

The assessment processes have an important role in the proposed methodology for developing hierarchical monitoring and assessment models (Chapter 4). In the dissertation, they are supported by the advanced data analysis techniques.

2.3 Assessment in Monitoring Systems

In general, the new data analysis techniques implemented in BI 2.0, such as text mining, proactive alerts, and predictive analytics, more often provide assessments than clear values as simple indicators. For example, even text mining techniques, such as text clustering and entity extraction, assess the similarity between texts and the word meaning.

In health care, monitoring systems are focused on the health care network and business processes as well as on individual patients. The following are two examples of the monitoring of individual patients based on the assessments. The first is a system for the monitoring of older adults at home in order to improve their health care (Zouba, et al., 2009). The system monitors their behavior, looks for changes, and automatically recognizes critical situations. The input data is gathered by video cameras and environmental sensors. The assessment of the situations is based on techniques of video analysis, sensor analysis, and activity recognition that combine video and environmental events to recognize complex activities (alarm states).

Another example of a monitoring system based on assessments is a cell-phone-based real-time monitoring system focused on cardiovascular disease (Zhanpeng, et al., 2009). The system is capable of performing continuous on-line electrocardiogram processing, automatically detecting and classifying abnormal cardiovascular disease conditions, and generating personalized cardiac health summary reports. The system is composed of various body-sensors providing data for the model that assess the particular state of individual conditions. The assessment of the disease condition is based on adaptive artificial neural network techniques.

On the other hand, the monitoring systems are focused on various aspects of the health care networks. In general, these systems are characterized by standardized indicators and OLAP techniques. This dissertation describes also the usage of advanced data analysis techniques for the assessment of various aspects of the health care systems.

2.4 Monitoring of Health Care Networks

A national Health Care Network (HCN) is a complex system composed of various Health Care Providers (HCP) that daily produce large quantities of data (Haux, 2006; Babulak, 2006). These data collected by medical institutions and governmental public health institutions can serve as valuable sources of evidence that should be taken into account when deciding about priorities to be included into strategic plans, or when deciding about specific measures to solve a particular health care problem. Thus, they can be periodically or continuously analyzed in order to provide government officials, development managers and civil society with information for improving internal processes, planning (Williams, 1999), and resource allocation.

The Slovenian national HCN is composed of different HCPs that are divided into the primary, secondary, and tertiary health care levels. The Primary Health Care Network (PHCN) consists of four sub-systems: general practice, gynaecology, pediatrics, and dentistry. The secondary level consists of hospitals and specialist care providers. The tertiary level includes the University Medical Centre and other university departments. According to the national health care program (MoH, 2000), municipalities are appointed to manage PHCN on their territory according to the model that must be prescribed by the Slovenian Ministry of Health (MoH).

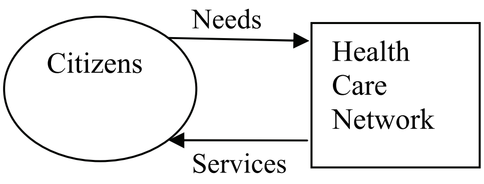


Figure 1: Health Care Network as a public service.

The monitoring of the current state and assessment of the future needs of the HCN services are requirements for the management of the network (Figure 1). In general, the prediction of citizens' needs is based on demographic data, environmental impacts, and the current state of health of the citizens. The managers have to allocate the capacity of the HCP in accordance with the predicted needs. E.g., the prediction system can assess how many patients will catch a cold in wintertime, so that managers can allocate health care resources in accordance with the predicted needs. Considering that predictions are not always reliable, the managers have to monitor the current state of the HCN. They need a monitoring system that provides information about health treatments, citizens' health conditions, the capacity of health care providers, and the efficiency of health care processes. Thus, the monitoring and assessment system for real-time monitoring of current state and predicting of future needs is necessary for the management of the HCN.

This monitoring system also has to be able to identify anomalies that can worsen the HCN, such as health care-deprived groups and areas (Section 5.1), populations with poor health conditions, increased expenses of HCN, physicians without licenses or formal education (Section 5.2), and other expected and unexpected anomalies related to HCN.

Considering some characteristics of the HCN, the monitoring of these systems is not an easy task, particularly if we are focused on the monitoring of small areas and many different providers. For example, Slovenia has about 2 million citizens, and it is divided in more than 210 municipalities, of which some have fewer than 1000 residents and some of these do not even have a health care provider. Hence, many patients visit HCPs in other municipalities. Thus, the measurement the PHCN availability for the patients from a municipality cannot be based on the simple ratio of the number of physicians and the total number of listed patients from the municipality (Section 5.65.6). Another problem is that not all physicians work full time and some of them work in two or three different locations. Moreover, many physicians provide services in various fields. All these characteristics request a specific methodology to the design of the monitoring model.

The HCN is characterized by many aspects, each one of which can be monitored by various indicators. For example, the aspect “physical accessibility of health care providers for citizens” can be assessed by indicators such as the road or air distance to the nearest health care provider available 24 hours a day or the nearest providers available at least once a week (Section 5.4). In order to improve the comprehensibility of the model, the proposed methodology for developing the HAMM introduces the concept of a module and assumes that all indicators aimed at the same aspect are grouped in the same module.

Many indicators are based on representative households or key informant surveys using either face-to-face or postal interviews. In the proposed PHCN monitoring and assessment model (Section 5), the analyses are based on data already provided by standard procedures. This makes a considerable difference in the type, quality, and quantity of collected input data, affects the selection of indicators, and justifies the different methodological approach taken in our case.

Many frameworks for the HCS assessment include indicators composed of sub-indicators. These composed indicators are usually calculated as a sum of the values of sub-indicators multiplied by weights, such as the Health Consumer Index (HCP, 2007), responsiveness (WHO, 2000), and overall assessment of the HCS. However, this is not always the best solution. For example, considering the assessment framework proposed by WHO (WHO, 2000), an HCS with a high population health, low responsiveness and bad fairness in financial contribution has the same overall assessment as the HCS with a low population health, high responsiveness, and good fairness. In this case, the different HCS have the same overall assessment. To avoid such problems, the indicators in the proposed PHCN monitoring model are combined in other ways, using multidimensional graphs, pivot tables, techniques of multi-criteria decision models, and the DM methods in the sense of discovering patterns in the indicators. For example, unusual values of performance indicators are detected by discovering small clusters.

3 Advanced Data Analysis Techniques for Monitoring Systems

This chapter describes some advanced data analysis techniques that can be implemented in monitoring systems. These techniques can provide additional information related to the monitored systems, such as the meaning of the measured values for the past, present, and future. We classify these techniques into the assessment models, data mining techniques, and advanced data visualization techniques.

3.1 Assessment Models

In general, these models join input parameters and assess various alternatives, situations, relations, scenarios, possibilities, etc. There are two approaches to the development of assessment models.

The first one is based on the knowledge acquired directly from experts that is implemented into the assessment model. Such an approach is utilized in Analytical Hierarchical Process (AHP) (Saaty, 1994), techniques for creating composite indicators (OECD, 2008), and the design of multi-criteria assessment models (Bohanec, 2006). The AHP reduces complex decisions to a series of pair wise comparisons, and synthesizing the results. For the development of rule-based multi-criteria decision models, we were using DEXi (Bohanec, 2011).

On the other hand, the development of the models can be based on supervised learning techniques (Section 3.2.3). These approaches discover knowledge in the learning data set and put it into the model. In general, all DM techniques that generate rules (e.g., classification, association, clustering, and sequence), or formulas (e.g., regression) can use this methodology. The usage of these techniques in monitoring models is described in Section 7.2.

3.2 Data Mining Techniques

This section describes some DM techniques that can be used in monitoring systems. The DM techniques are defined as automated searching of large bodies of data for the extraction of new information and previously unrecognized relationships (Brown, et al., 2007). In general, various definitions stress that DM techniques are automated or semi-automated methods aimed at the discovery of non-trivial and unexpected information in large amounts of data. These characteristics make some DM techniques useful for the monitoring systems.

DM techniques can be classified into descriptive and predictive (Han and Kamber, 2006). While the former aim at finding human-interpretable patterns and associations, the latter search for strong patterns that can be generalized in order to predict interesting aspects of the system. The descriptive techniques include unsupervised learning techniques, where only input data is processed. These methods are:

- Clustering: finding and visually documenting groups of facts that have not been previously known.
- Association rules discovery: looking for patterns where one event is connected to another event.
- Sequences discovery: looking for patterns where one event leads to another later event.
- Text analyses: text clustering and data extraction.

Predictive mining methods include supervised learning techniques, where target outputs as well as the inputs are processed in order to construct a model that produces the correct output. These methods are:

- Classification: assigning or classifying data records into categories of a class variable. The class variable is categorical and may have several possible values.
- Regression: predicting the value of a numeric-dependent variable.

Some DM techniques that can be used in monitoring systems are briefly described in the next sections. The real usage of these techniques in monitoring models is described in Section 7.2.

3.2.1 Association Rules Discovery Techniques

The purpose of association rules discovery techniques is to find items in data that are associated with each other in a meaningful way (cause and consequence – antecedent and consequent) (Agrawal, et al., 1996). The association rules are composed of body and head. Rule body (*A*) includes one or more items, which imply the presence of another item. Rule head (*B*) includes an item whose presence is inferred from the items in the rule body. The association rules are described by parameters *Support*, *Confidence*, and *Lift,* which measure their strength.

Support is the percentage of all data records containing both the body (*A*) and the head (*B*).

$$Support = \frac{|A \cap B|}{|S|} \tag{1}$$

Here, the notation |*X*| represents the size of the set *X*, and |*S*| represents the total number of data records.

Confidence is defined as the likelihood that the head item will be in the record, given the presence of the body item(s). It is the ratio of records containing all of the items in (*A*) and (*B*) to records containing the items in (*A*).

$$Confidence = \frac{|A \cap B|}{|A|} \tag{2}$$

Lift is the degree to which the Confidence is larger (or smaller) than expected. It is defined as the ratio between Confidence and expected confidence. Expected confidence is the ratio of records having the consequent items (B) and the total number of records.

$$Lift = \frac{|A \cap B| / |A|}{|B| / |S|} \tag{3}$$

3.2.2 Clustering

The general purpose of clustering is to discover knowledge in data by grouping similar objects (data items) into same groups. Each group, called a cluster, consists of objects that are in some way similar to each other and dissimilar to objects in other clusters. Clustering brings comprehensibility that can help reveal new useful information in the data.

Clustering is characterized by large datasets with many attributes, and many clustering techniques. They became popular by intense developments in information retrieval and text mining (Steinbach, et al., 2000; Dhillon, et al., 2001), spatial data analysis (Ester, et al., 2000), sequence and heterogeneous data analysis (Cadez, et al., 2001), Web applications (Foss, et al., 2001; Heer and Chi, 2001), DNA analysis in computational biology (Ben-Dor and Yakhini, 1999), anomaly detection (Patcha and Park, 2007; Münz, et al., 2007), and many others (Berkhin, 2006).

Clustering methods may be classified into partitioning, hierarchical, density, and grid-based methods (Han, et al., 2001). Alternatively, they can be classified into partitioning, hierarchical, density-based, grid-based, and model-based methods (Nathiya, et al., 2010). Berkhin (2006) also provides a detailed classification of the clustering methods. He classified the methods into hierarchical methods, partitioning methods, grid-based methods, constraint-based clustering, methods based on co-occurrences of categorical data, clustering algorithms used in machine learning, scalable clustering algorithms, and algorithms for high dimensional data clustering.

Hierarchical clustering is subdivided into agglomerative and divisive. An agglomerative method starts with each point as a separate cluster, and successively performs merging until a stopping criterion is met. A divisive method begins with all points in a single cluster and performs splitting until a stopping criterion is met. The result of a hierarchical clustering method is a tree of clusters called a dendogram.

While hierarchical algorithms gradually (dis)assemble points into clusters, partitioning algorithms learn clusters directly. In doing so they try to discover clusters either by iteratively relocating points between subsets (Partitioning Relocation Clustering), or by identifying areas heavily populated with data (Density-based Partitioning). The partitioning methods start with some arbitrary initial clusters and iteratively reallocate points to clusters until a stopping criterion is met. Density-based clustering methods try to find clusters based on the density of points in regions. Dense regions that are reachable from each other are merged to formed clusters. The grid-based methods work with data indirectly by constructing summaries of data over the attribute space subsets. They perform space segmentation and then aggregate appropriate segments. Grid-based methods are fast and handle outliers well. The point-by-attribute representation for categorical data is high dimensional and extremely sparse. In this situation, conventional clustering methods, based on similarity measures, do not work well. The idea of categorical data Co-occurrence comes to the rescue. The algorithms such as ROCK, SNN, and CACTUS are methods based on the Co-occurrence of Categorical Data (Berkhin, 2006). The clustering of large datasets is based on Scalable clustering algorithms. The survey of clustering algorithms by Xu and Wunsch (2005) shows that no one clustering algorithm can be universally used.

Clustering techniques can be used in monitoring models for the identification of groups of customers, outlier detections for identification of fraudulent events, pattern recognitions for monitoring of human behaviors, automatic grouping of document in accordance with contents, etc. We identify the following general requests for the clustering algorithms used in monitoring systems:

- handling a large volume of numerical and nominal data as well as high-dimensional features with acceptable time and storage complexities,
- detecting or removing possible outliers and noise,
- fast provision of results,
- decreasing the reliance of algorithms on users-dependent parameters, and
- good data visualizations and provision of results without deep knowledge of clustering methods.

3.2.3 Predictive Data Mining Techniques

Some of the best-known predictive data mining algorithms are (Wu, et al., 2007):

- C4.5,
- Support vector machines (SVM),
- Naive Bayes,
- AdaBoost,
- K-nearest neighbor,
- Classification and Regression Trees (CART), and
- Artificial Neural Networks.

C4.5 (Quinlan, 1993) is a descendant of CLS (Hunt, et al., 1966) and ID3 (Quinlan, 1979). It was superseded in 1997 by the commercial system See5/C5.0 (or C5.0 for short). Support vector machines (SVM) (Vapnik, 1995) are considered to be among the most robust and accurate classification algorithms. Another very important algorithm is the Naive Bayes, which is very easy to construct, and does not need any complicated iterative parameter estimation schemes. The AdaBoost algorithm (Freund and Schapire, 1997) is one of the most important ensemble methods (Dietterich, 1997), which employ multiple learners to solve a problem. The k-nearest neighbor classification (Fix and Hodges, 1951; Tan, et al., 2006) finds a group of *k* objects in the training set that are closest to the test object, and bases the assignment of a label on the predominance of a particular class in this neighborhood. The Classification and Regression Trees (CART) (Breiman, et al., 1984) represent a major milestone in the evolution of Artificial Intelligence, Machine Learning, non-parametric statistics, and data mining. A CART tree is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample. This learning technique produces either classification or regression trees, depending on whether the dependent variable is categorical or numerical. The Artificial Neural Networks are based on theoretical foundations of neural networks in the brain, which are composed of neurons and where the weights of connections between neurons are adjusted on the basis of learning data (Widrow, et al., 1994).

Prediction is a two-step procedure. In the first step, a model is trained by using a training sample. Each sample is described by attributes where the output attribute values indicate the predefined class or value to which the sample belongs. The output attribute can be discrete or continuous. The prediction of the discrete attributes is based on the classification algorithms. The prediction of continuous attributes is based on regression techniques. In the second step, the model attempts to classify the objects, which do not belong to the training sample and form the validation sample. This process, which is known as supervised learning, can be used for the creation of assessment models, e.g., for prediction of the future states of the monitored systems.

3.3 Advanced Data Visualization Techniques for Monitoring Systems

Data visualizations (Keim, et al., 2008) utilize human visual perceptual capability in order to enhance information comprehension (Friendly, 2006). This section describes advanced data visualization techniques suitable for use in monitoring systems. In the dissertation, these techniques are aimed at the comprehensible visualization of results from various perspectives in order to reveal their meaning. For instance, the environmental monitoring systems can show various indicators on maps (Ladner, et al., 2009). Thus, map areas are highlighted from perspectives such as the weather conditions, biotic diversity, and pollution.

In striving for a better understanding of information visualization, various classification schemes have been proposed over the past few years. Shneiderman (1996) suggested taxonomy for information visualization designs built on data type and task, the type by task taxonomy. He distinguished the following data types:

- 1-dimensional: text files and alphanumeric list of names,
- 2-dimensional: geographic map or book layout,
- 3-dimensional: real world objects and chemical molecules,
- temporal: time-series and scientific measurement rows,
- multi-dimensional: relational database content,
- tree: structured data collections with hierarchy constraints, and
- network: structured object sets that do not conform to tree constraints.

Shneiderman (1996) considered that his classification was incomplete and forecasted upcoming applications would require novel and, respectively, specialized data structures.

Keim (2003) concentrates on the design of the visual environment and suggests a classification of visualization designs that takes into consideration the following dimensions:

- data type to be visualized,
- visualization techniques, and
- interaction technique.

Usually, it is assumed that these dimensions are orthogonal. Orthogonality means that any of the visualization techniques may be used in conjunction with any of the interaction techniques for any data type.

According to Keim (2003), data types include one-dimensional data, such as temporal data; two-dimensional data, such as geographical maps; multidimensional data; text and hypertext; hierarchies and graphs, such as telephone calls; and algorithms and software.

There are a number of well-known visualization techniques, such as X-Y plots, line plots, and histograms. These techniques are useful for data exploration but are limited to relatively small and low-dimensional data sets. The visualization techniques can be classified into the following groups Keim (2003):

- Standard 2D/3D displays: the techniques such as bar charts, X-Y (X-Y-Z) plots, line graphs, maps, and scatter plots deploy traditional visual encoding. These techniques are useful for data exploration but are limited to small and low-dimensional data sets.
- Geometrically transformed displays: these techniques aim at finding interesting transformations of multidimensional data sets. This group of techniques includes landscapes, hyperbolic plans, scatter plot matrices (Cleveland, 1993), and parallel coordinates (Inselberg and Dimsdale, 1990).

- Icon-based displays: the idea in this technique is to map the attribute values of a multi-dimensional data item to the features of an icon. The icons can be Chernoff faces (Astel, et al., 2006), stick figures, needle icons, heat maps, and star icons.
- Dense pixel displays: the basic idea of pixel-oriented techniques is to map each data value to a colored pixel and present the data values belonging to one dimension (attribute) in a separate sub window. The value ranges of the attribute are mapped to a fixed color map (Keim, et al., 2003).
- Stacked displays: the idea here is to present data portioned in a hierarchical fashion. This group includes techniques such as treemaps (Shneiderman, 1992) or dimensional stacking (Langton, et al., 2007). The treemap is a popular method for visualizing hierarchical data. By dividing the display area into rectangles recursively according to the hierarchy structure and a user-selected data attribute, a treemap can effectively display the overall hierarchy as well as the detailed attribute values.

Interaction techniques allow users to interact with the visualizations. They may be classified into interactive projection, filtering, zooming, distortion, linking, and brushing.

4 Methodology for Developing Hierarchical Assessment and Monitoring Model

In order to improve the functionality and comprehensibility of monitoring systems, we propose a new methodology for developing monitoring models. The methodology is based on the following three key concepts:

- Hierarchical assessment and monitoring models (HAMM) (Section 4.1),
- Generic Templates (GTs) (Section 4.2), and
- Advanced data analysis techniques (Chapter 7).

The HAMM is a monitoring model (MM) that shows the basic concept, all details, and relations between various aspects of the monitored system. Some concepts of these MMs are generic. These concepts can be used for monitoring of similar systems in various fields. The usage of these concepts can simplify the development of monitoring models. We call these concepts GTs. These concepts are modeled by the HAMM.

This methodology also includes advanced data analysis techniques. In principle, any data visualization method can be used, such as pivot tables, charts, network graphs, and maps. The same holds for data analysis methods, which can include Structured Query Language (SQL) procedures, Online Analytical Processing (OLAP) techniques for interactive knowledge discovery, as well as knowledge discovery in databases (KDD) and DM techniques. Thus, these models are aimed at simple monitoring as well as assessments and predictions.

The details of the HAMM and GT are described in the next sections.

4.1 Hierarchical Assessment and Monitoring Models

Despite the many frameworks related to the monitoring of various systems, such as Data-driven Decision Support System (DSS) (Power, 2002), Performance Monitoring, Business Performance Management (BPM), Business Activity Monitoring (BAM) (Dresner, 2003) etc., there is a lack of methodologies for describing the concept and details of the monitoring models. This is particularly true when the monitoring systems include various data analysis and data visualization techniques.

Thus, we propose a new methodology based on the model that we call HAMM. The model is composed of hierarchically structured and connected modules. Each module is aimed at monitoring of a particular aspect of the monitored system, which is of interest for decision-makers and managers. The module includes at least one monitoring process (MP) that is aimed at a particular aspect of the monitored system. All MPs in the module show the same aspect of the monitored system from a slightly different perspective and using various technologies. The structure of the connected modules shows the basic concept of MM. The descriptions of MPs show the details of MM.

In this dissertation, the term MP is used in a broad sense. It denotes a periodically or continuously performed data analysis process without distinguishing between “monitoring,” “validation,” and “assessment.” Notice that the term MP has a wider meaning than the term “indicator” because it covers a whole process of particular data analyses. Each MP is characterized by:

- input data,
- data analysis methods,
- output data,
- output data presentation methods,
- monitoring objectives,
- target values,
- data collection methods,

- constraints on the data,
- security requirements,
- data dimensions, and
- possible side effects of MPs.

Among these dimensions, the data analysis methods transform the input data to output data represented by some output data presentation methods according to the given monitoring objectives. Usually, the output data consist of different indicators presented by various data visualization techniques. The term "data visualization" is used in its broad sense. It includes – mostly visual – presentations of both single data elements and complex patterns, and it does not explicitly distinguish between "data," "information," "pattern," and "knowledge."

The target level is a level of performance that the organization aims to achieve for a particular activity. With respect to organizational goals, the output data in the model are classified into the lead and the lag (Niven, 2003). The lead ones show the performances that have an influence on achieving the goals, whereas the lag ones are related to the degree of achieving the goals. Information about data collection defines how and how often the data has been collected or needs to be collected. For example, data can be collected by representative surveys or by standard procedures in organizations according to some refreshment rate. The constraints on the data define the valid input and output data. The security requirements define the use and management of the monitoring processes and of the data.

Our monitoring models also address the possibility of side effects that may be induced as a reaction to the monitoring. Such an unwanted consequence is the so-called "perverse" learning that happens when organizations or individuals learn which aspects of performance are measured and which are not, so they use this information to manipulate their assessments. For example, among cardiac surgeons in New York whose individual unadjusted patient death rates have been published regularly, there has been a tendency to avoid taking on high-risk cases with a subsequent increase in mortality of patients at risk for cardiac surgery (Dranove, et al., 2002). The unintended consequences could also be caused by detailed monitoring of individuals; the monitoring system that severely limits individual freedom could be counter-productive.

In order to improve the comprehensibility of model, the modules aimed at certain aspects of a monitored system are hierarchically organized. The MPs in higher-level modules are hierarchically connected with MPs at lower levels in the hierarchy. Each connection is a data channel that connects the outputs of lower-level MPs with the inputs of higher-level MPs. In principle, each connection denotes that the output data of lower-level MPs can help to explain the output data of MPs at higher levels. The lowest-level MPs provide independent variables that can be manipulated or changed. A suitable user interface enables the users to move from data presented by higher-level MPs to appropriate data presentation of lower-level MPs and vice versa.

The proposed methodology is not limited to any particular data presentation or data analysis method. In principle, any data presentation method can be used, such as pivot tables, charts, network graphs, and maps. The same holds for data analysis methods, which can include Structured Query Language (SQL) procedures, Online Analytical Processing (OLAP) techniques for interactive knowledge discovering, as well as knowledge discovery in data bases (KDD) and data mining methods for discovering important but previously unknown knowledge.

The process of HAMM development is divided into the following phases:

1. Definition of the purpose and objectives of the monitoring system, and relevant available data.
2. Definition of all interesting aspects – modules – of the monitored system, and their relations. These definitions and relations can be based on a suitable GT. In this phase, a basic concept of the HAMM is designed.
3. Design of indicators for monitoring of interesting aspects of the monitored system.
4. Definition of MPs implemented in the modules. In this phase, the advanced data analysis techniques for assessments of the indicated values are implemented.
5. Validation of the developed model.

4.2 Generic Templates

Some basic monitoring concepts are generic; they can be used for the monitoring of similar systems in various fields. For example, public services systems such as health care, police, fire brigade, and an educational system have many common elements and processes. At the top level, they all include service providers, services, and citizens using these services. All these systems have to provide equal and adequate levels of services for all citizens.

Therefore, the proposed methodology includes GTs, i.e., templates for MM aimed at systems that are similar in structure and functionality. A GT describes the basic concepts of the MM composed of hierarchically connected modules, but without detailed descriptions. GTs can be used as a starting point in order to simplify the design and improve the comprehensibility of the monitoring models. The following three sections describe three typical examples of GTs.

4.2.1 GT for Monitoring and Assessment of Public Service Availability and Resource Allocation

Public service is a term usually denoting services provided by governments to their citizens, such as health care, police, fire brigade, public transportation, an educational system, social services, and so on. In general, the systems providing these services are composed of providers covering certain areas, their services, and the populations who need these services. These systems have to provide an adequate and equal level of services for all citizens. If providers in a certain area are overloaded then citizens in this area will experience a shortage of services. On the other hand, if the provider's resources are oversized then the expenses will increase. Thus, the management of these systems requires careful and timely monitoring.

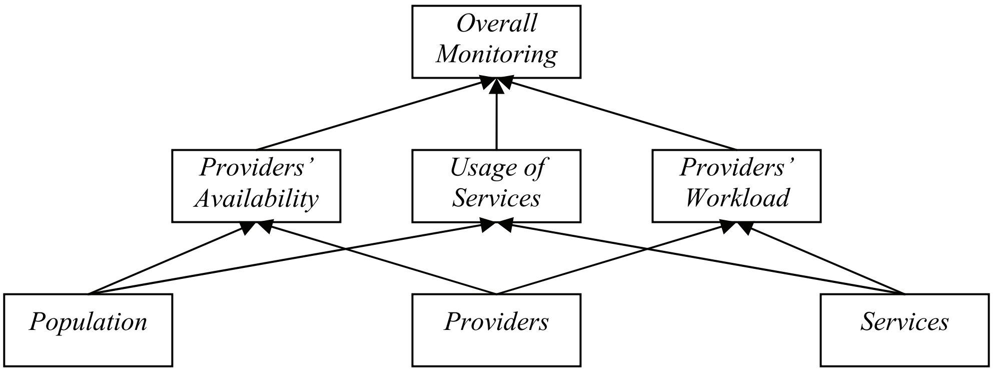


Figure 2: GT for monitoring the availability and resource allocation of public services. Adapted from MediNet++ (2008).

The GT for Monitoring and Assessment of Public Service Availability and Resource Allocation is shown in Figure 2. The template is composed of seven hierarchically connected monitoring modules. The modules *Population*, S*ervices*, and *Providers* are at the lowest level (Figure 2), and provide aspects about the population that needs services, about the activities of service providers, and about the providers who offer these services.

Different combinations of the module output data provide new aspects of the public system. For example, the aspect provided by the module *Providers' Workload* is based on the combination of data from the modules *Services* and *Providers*. This module is aimed at the workload of the service providers. It provides information such as the number of service activities per provider considering the time needed for the services and provider's capacities.

The module *Providers' Availability* refers to the providers' availability for the population. The module provides information such as the number of service providers per citizen considering the population's needs and provider's capacities.

The module *Usage of Services* shows the real use of the provider's services. It includes data from the modules *Services* and *Populations*. This module provides information such as the number of used services per citizen. This module shows the population's needs, while at the same time detects anomalies such as areas or groups of people having higher rates of required services.

The output data from the second level modules are combined into the topmost module *Overall Monitoring*. This module provides an overall picture of the monitored system. The proposed PHCN monitoring model presented in Chapter 5 is based on this GT.

4.2.2 GT for Maintenance Monitoring

The Enterprise Asset Management (EAM) means the whole-life optimal management of the physical assets of an organization to maximize value. It covers the design, construction, commissioning, operations, and maintenance and decommissioning/replacement of plant, equipment, and facilities. The important part of the EAM is a system for assets maintenance. This section is focused on the GT aimed at the monitoring of the maintenance system (Figure 3). The monitoring is particularly important for condition-based maintenance where the prevention is based on the monitoring of the physical variables that determine the symptoms of a failure (Hao, et al., 2010).

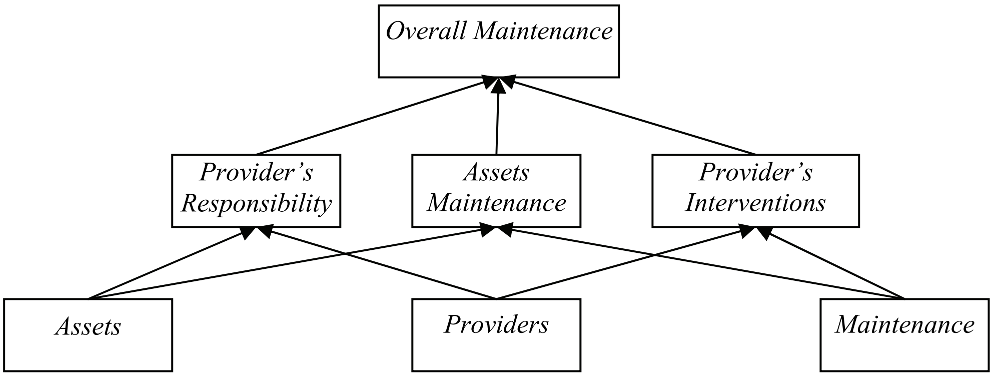


Figure 3: GT for monitoring asset maintenance systems.

In general, maintenance systems in various organizations have a similar structure and functionality. Thus, the monitoring systems aimed at the maintenance can be described by the same GT, which we proposed in Figure 3. The template is composed of seven hierarchically connected modules – aspects. The basic low-level aspects are *Assets*, *Providers*, and *Maintenance*. They provide basic aspects about assets in organizations, about providers who repair and upgrade the assets, and about the maintenance of the assets.

The combinations of output data from basic modules provide new aspects of the asset management system in organizations. For example, the results provided by the module *Provider's Interventions* based on the combination of data from the modules *Maintenance* and *Providers* are aimed at provider's activities. The module provides information such as the number of repairs per providers, and response time of providers.

The module *Providers' Responsibility* refers to monitoring of the assets that have to be repaired and upgraded by the provider who is responsible for them. The module is based on the combination of data from the modules *Assets* and *Providers*. The module provides information such as who is responsible for the maintenance of critical assets in the company, and the share of assets that are the responsibility of certain service providers.

The module *Assets Maintenance* is based on the combination of data from the modules *Assets* and *Maintenance*. It monitors the repairs and upgrades of the assets, and it provides information such as the availability of a certain asset, the maintenance costs, and preventive maintenance of the asset.

The output data from the second-level modules are combined in the module *Overall Maintenance*. This module provides the main information about the asset management in the organization.

4.2.3 GT for User Activity Monitoring

Usually monitoring systems are aimed at the monitoring of various systems such as business organizations, and public services. In this section, we propose a GT that is focused on the monitoring of current users' activities in order to provide the additional and useful information related to these activities. Thus, this system assesses which additional and available information can be interesting for users considering their activities.

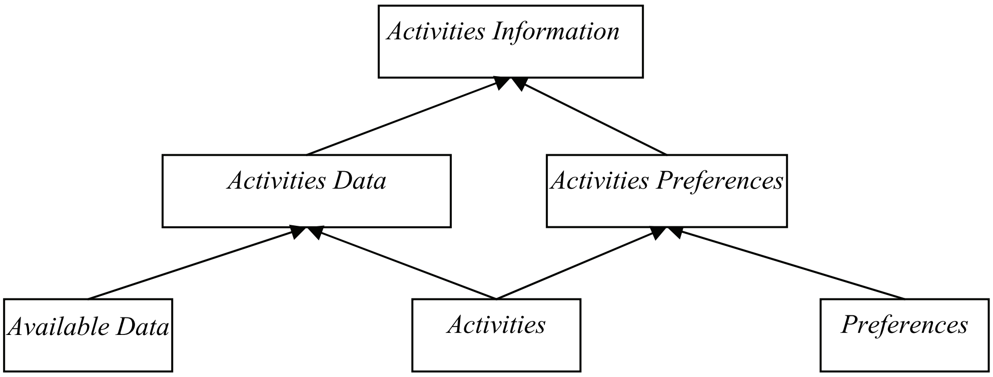


Figure 4: GT aimed at the provision of useful information related to current activities.

This GT is composed of six modules, which are shown in Figure 4. The low-level modules provide aspects on the available data, the current user activities, and the user preferences. Thus, the module *Available Data* is aimed at the public and internal data available in various electronic formats. The module *Activities* provides information about current user activities such as text that a user currently reads or writes; information provided by voice, speaker, face, fingerprints, or license plate recognition systems; geographic information provided by GPS; and so on. The module *Preferences* provides information about a user's fields of interest, known people, towns visited in the past, and so on.

On the second level, there are two modules (Figure 4). The module *Activities Data* is focused on available data that are related to the user's current activities. For example, this module provides information that already exists in a company and it can help a user to deal with current problems. In a similar way, the module *Activities Preferences* provides useful information that is related with data in the module *Preferences*. For example, when users read a mail, this system can automatically show that some people mentioned in the mail are also in their address book. The top-level module *Activities Information* provides the most useful information related to the user's current activities. The purpose of this module is to help users with the additional information to solve current problems. This GT is interesting for our further work because it deals with structured and unstructured data.

5 Primary Health Care Network Monitoring Model

In accordance with the proposed methodology for developing monitoring models (Chapter 4), we developed a model for the monitoring of resources in the PHCN. The model is based on the GT for the monitoring and assessment of public service availability and resource allocation (Section 4.2.1). The aim of the PHCN monitoring model is to improve decisions related to PHCN planning, which should be focused on the appropriate allocation of health care resources to those who need it in accordance with the national strategy. The model includes the following hierarchically connected modules:

- The *PHCN Assessment* module addresses the main aspect of the PHCN, which helps to identify inequalities and other anomalies related to the allocation of health care resources. This aspect presents general information about the PHCN. Thus, the monitoring processes in this module are designed for top managers rather than middle or low-level managers.
- The *Visits* module is aimed at the rate of patients' visits to PHCN. This aspect is important for the assessment of patient's needs, and it shows some anomalies related to a level of population health. For example, an increased number of visits overloads the HCPs, and requires greater health care capacity. On the other hand, an increased number of visits can identify a low level of sub-population health caused by unhealthy living conditions.
- The *Accessibility* module is aimed at the physical accessibility of patients to HCPs. In accordance with the national program of health care, all population in Slovenia must live within an acceptable distance to the nearest HCP. Thus, the monitoring of the PHCN from this aspect is mandatory.
- The *Unlisted Patients* module is aimed at the ratio between listed patients and overall residents in a given area. At the primary level, the residents are usually listed by their physicians. This module shows some health care-deprived areas. For example, the population from a certain area rarely visits the HCPs, and usually individuals do not have their own physicians. The deprivation in these areas can be overlooked, because the other aspects such as *PHCN Availability* take into account only the listed patients (see MP F2 and MP F4 in Section 4). A high number of unlisted patients can also result from a low quality of HCPs. Thus, in our model, the monitoring of this aspect is mandatory.
- The *PHCN Availability* assesses the availability of PHCN for patients from a certain area considering the health care capacity and needs of the listed patients. Again, the listed patients are those who have their own physicians. Thus, this module provides information about the PHCN availability for the listed patients from certain areas. This is different from the *HCP Availability* module, which shows the availability of each HCP or physicians for their own patients. In accordance with the national program of health care, equal availability of PHCN is required, thus the monitoring of this aspect is mandatory, too.
- The *Physicians* module shows the human resources included in the PHCN. This aspect is very important, because the physicians and nurses who are working in the PHCN are the most important part of this system. A PHCN manager has to know where more (or less) physicians or nurses are needed now or in the future, particularly if we consider that health studies take many years to complete. The manager also needs to know the education level and obtained licenses of individuals who work in the PHCN, because in the past some employed physicians were discovered not to have a formal qualification for their job.

- The *HCP Availability* module is aimed at the assessment of the availability of HCPs or physicians for their patients considering the provider's capacity and the patients' needs. Thus, this module shows the discrepancy in availability of different HCPs. The information provided by this module could be interesting to HCP managers who want to compare their workload with other providers.
- The *Dispersion* module addresses the problem that some physicians work in more than one location. This dispersion of physicians usually means additional workload for them and their lower availability for patients at one of the locations. This module is related to the module *Physicians*.
- The *Health Activities* module provides basic data about health treatments.
- The *Population* module provides basic demographic data about the population living in a given area such as age, gender, municipality where they live, etc.
- The *Health Resources* module provides basic data about HCPs (e.g., number of physicians and locations of providers) and data about physicians (e.g., age, gender, and working time).

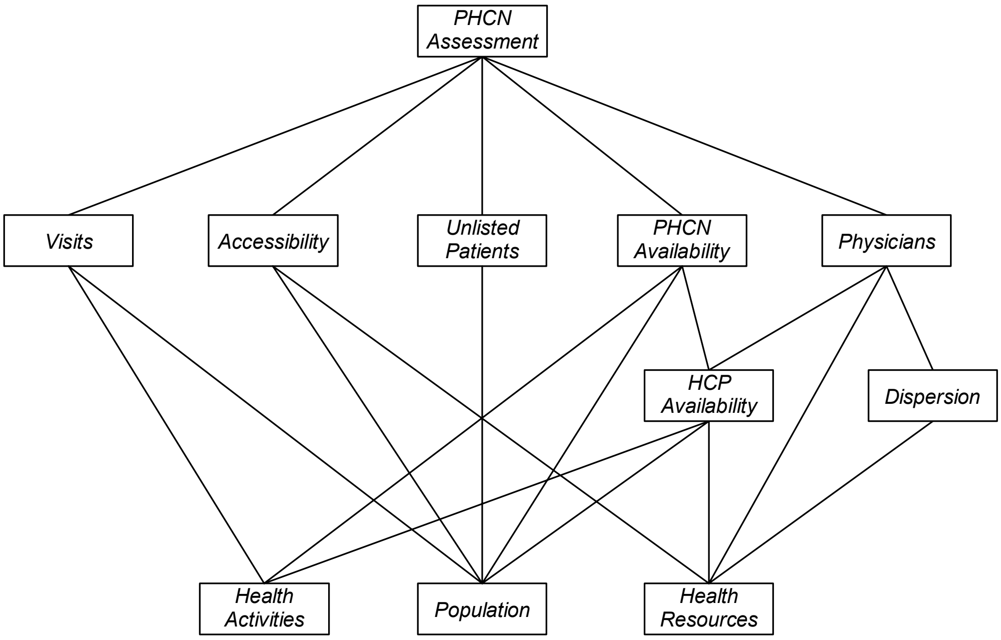


Figure 5: The basic concept of the PHCN monitoring model.

A more detailed hierarchy including modules and monitoring processes is shown in Figure 6. Each MP is denoted by a letter and number. The letter indicates the module that contains the MP, and the number indicates the index of the MP in the module. Some MPs provide output data in the form of named variables; in this case, abbreviated names of these variables are also shown in the corresponding MP rectangles.

Methodologically, such models are developed according to the phases as described in section 4.1. The structure can be developed in two ways: *top-down* by decomposition from general into more detailed modules, and *bottom-up* by the aggregation of modules. Effectively in practice, the development most often combines both, which is known as the *middle-out* strategy (Coiera, 2009; Bohanec, 2011).

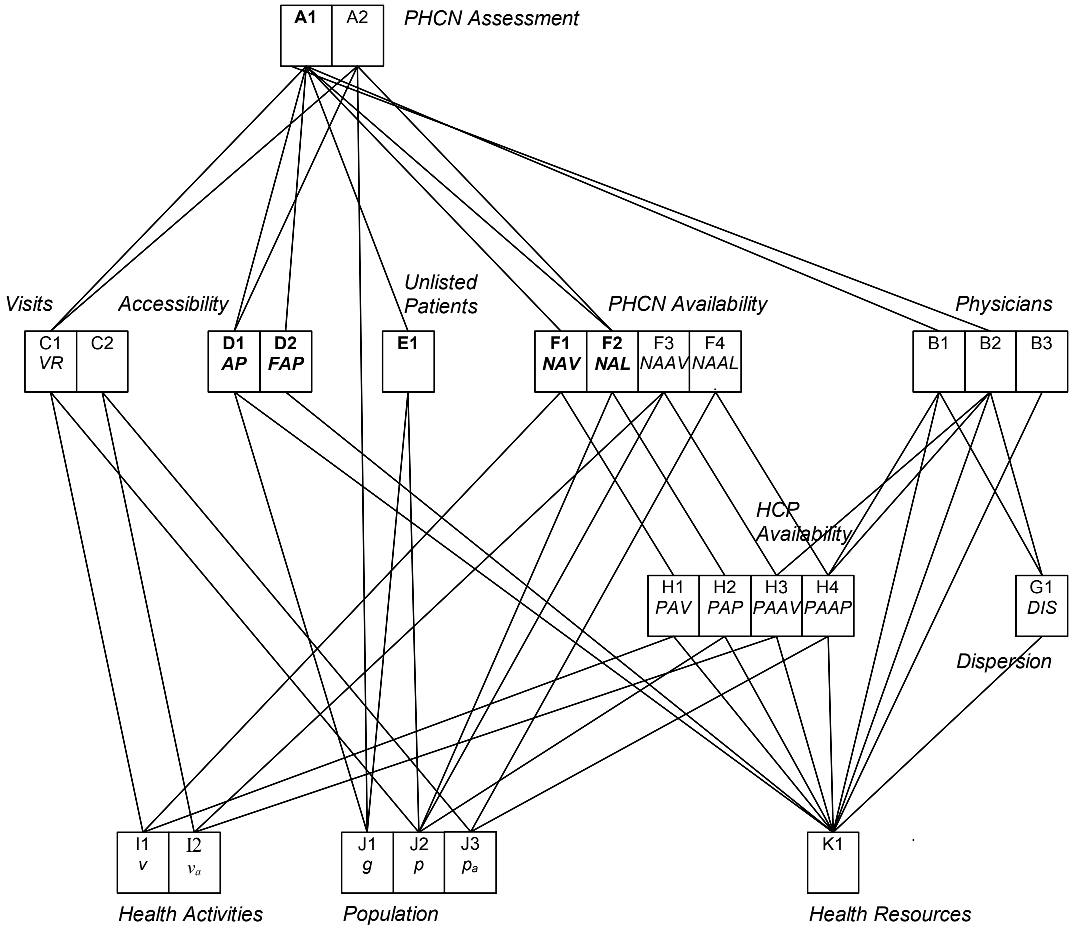


Figure 6: HAMM aimed at allocation of PHCN resources.

5.1 PHCN Assessment

The top-level module *PHCN Assessment* addresses the main aspect of the PHCN that helps to identify inequalities and other anomalies related to the allocation of health care resources. This aspect presents general information rather than a detailed explanation of anomalies, and the monitoring processes in this module are designed for top managers rather than middle or low-level managers. Detailed explanations of particular aspects shown by this module are provided by lower-level modules. Thus, the MPs in the *PHCN Assessment* module are based on analyses of output data from lower-level MPs. In principle, the MPs in this module must be able to process several input parameters using different methods of KDD and techniques of multi-criteria decision models.

The module *PHCN Assessment* includes two MPs, A1 and A2. A1 presents the main aspects of the PHCN in the form of short clauses, such as “more (or fewer) physicians (in terms of FTE, Full Time Equivalent) are needed.” The clauses are automatically generated based on input data provided by the lower-level MPs and assessment model that compare input data with the prescribed target values. This rule-based assessment model is manually created.

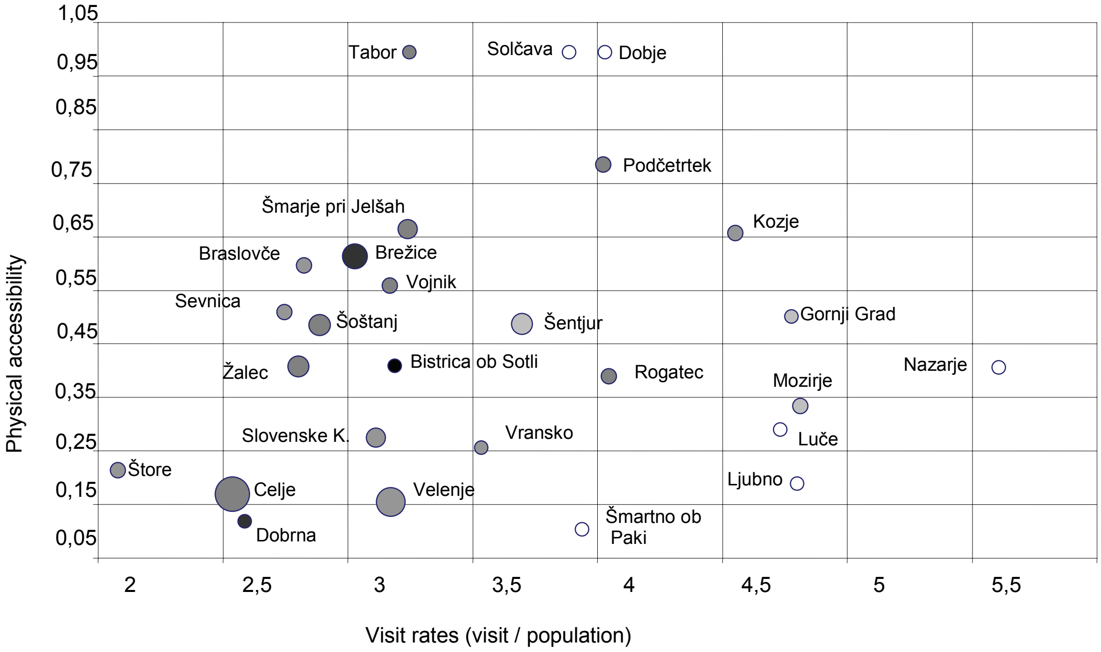


Figure 7: The overall view of the PHCN sub-segment of the general practice for the Celje region (year 2003). Physical accessibility is the proportion of the population that needs more time to get to the nearest HCP than is the maximal acceptable travel time, considering the length and the category of roads.

MP A2 presents the main aspects of the PHCN in municipalities by a multidimensional chart (Figure 7). Each municipality is represented by a dot, which displays four dimensions:

- The dot's horizontal position shows the visit rates of listed patients from municipalities (VR, see MP C1)
- The vertical coordinate corresponds to the physical accessibility of patients to providers (AP, see MP D1)
- The dot color intensity represents the network availability for visits (NAV, see MP F1)
- The dot diameter is proportional to the number of inhabitants in the municipality (*g*, see MP J1)

In this way, municipalities that have average values of VR and AP appear in the middle of the chart. The outliers represent more or less unusual municipalities with respect to particular aspects of PHCN. A more detailed and accurate explanation of reasons for such municipalities' positions in the chart is provided by the lower-level modules B–F (see Figure 6).

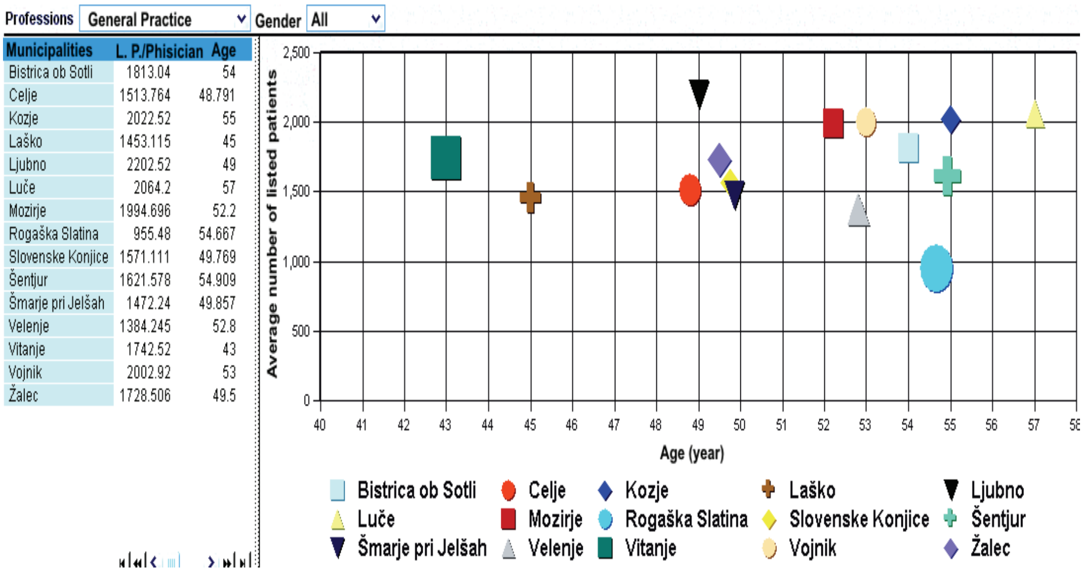
5.2 Physicians

The module *Physicians* shows the human resources included in the PHCN. This aspect is very important, because the physicians and nurses who are working in the PHCN are the most important part of this system. A PHCN manager has to know where more (or fewer) physicians or nurses are needed now or in the future, particularly if we consider that health studies take many years to complete. The manager also needs to know the education level and obtained licenses of individuals who work in the PHCN, because in the past some employed physicians were discovered not to have a formal qualification for their job.

The module *Physicians* intends to provide the overview of physicians in the PHCN. Thus, MP B1 presents the aspects of physicians characterized by their specialization, age, gender, workload, location where they work, and dispersion. The workload is inversely proportional to providers' availability for adjusted patients (PAAP), and dispersion means a number of locations where the physician works (see MP G1).

This MP is based on the OLAP technology (Figure 8). The scatter plot on the right-hand side of Figure 8 shows the average age, average workload, and average dispersion of physicians in municipalities for different specializations. The *x*-axis shows the average age of physicians, and the *y*-axis shows the workload measured by average number of listed patients per physician. Municipalities are shown by shapes and colors of data points as explained in the legend. The size of these points is proportional to the average dispersion of physicians in the municipality. The displayed physicians could be selected according to their specialization and gender using combo boxes in the top left corner. On the left-hand side, the average workload, and age in municipalities are represented by the pivot table (Figure 8).

This presentation clearly shows outliers and anomalies in the HCS. For example, the municipalities with workload and age of physicians above the average are shown in the top right corner. Detailed information about this interesting aspect can be found in the lower-level modules.



Professions: General Practice; Gender: All

Municipalities	L. P./Phisician	Age
Bistrica ob Sotli	1813.04	54
Celje	1513.764	48.791
Kozje	2022.52	55
Laško	1453.115	45
Ljubno	2202.52	49
Luče	2064.2	57
Mozirje	1994.696	52.2
Rogaška Slatina	955.48	54.667
Slovenske Konjice	1571.111	49.769
Šentjur	1621.578	54.909
Šmarje pri Jelšah	1472.24	49.857
Velenje	1384.245	52.8
Vitanje	1742.52	43
Vojnik	2002.92	53
Žalec	1728.506	49.5

Figure 8: The overview of GP in the Celje region (year 2004). The scatter plot shows the average age, workload, and dispersion of physicians in municipalities. The average workload (number of listed patients per physician), and age of physicians in municipalities are represented by the pivot table on the left side.

This module also includes MP B2 aimed at discovering interesting relations in physicians' data by the association rules discovery method. These rules can be presented by tables such as Table 1. The rules show the characteristics of dentists in some municipalities. For example, the municipality Kranj is characterized by male dentists younger than 40 years that are underloaded (see Rule 4). These rules can also be presented as a computer-generated text, such as (Rule 4): "Six percent of all young male dentists (not older than 40 years) with low workload work in Kranj, this is 2 times higher than expected considering the national average. This rule is supported by 0.41% of all Slovenian dentists." The rules provided by MP B2 can be used in synergy with OLAP-based visualizations provided by MP B1, because some interesting relations omitted by OLAP techniques can be found by association rules and vice versa.

Rule		Support	Confidence	Lift
1	[Workload: high] ==>Rogaška Slatina	0.21%	8%	19.3
2	[Age: 60+]+[Workload: low]+ [Gender: M] ==>Maribor	0.62%	20%	2.4
3	[Age: 60+]+[Workload: low]+ [Gender: M] ==>Murska Sobota	0.31%	5%	2.5
4	[Age: to 40]+[Workload: low]+ [Gender: M] ==>Kranj	0.41%	6%	2.0

Table 1: Association rules showing relations between dentists and municipalities in which they work.

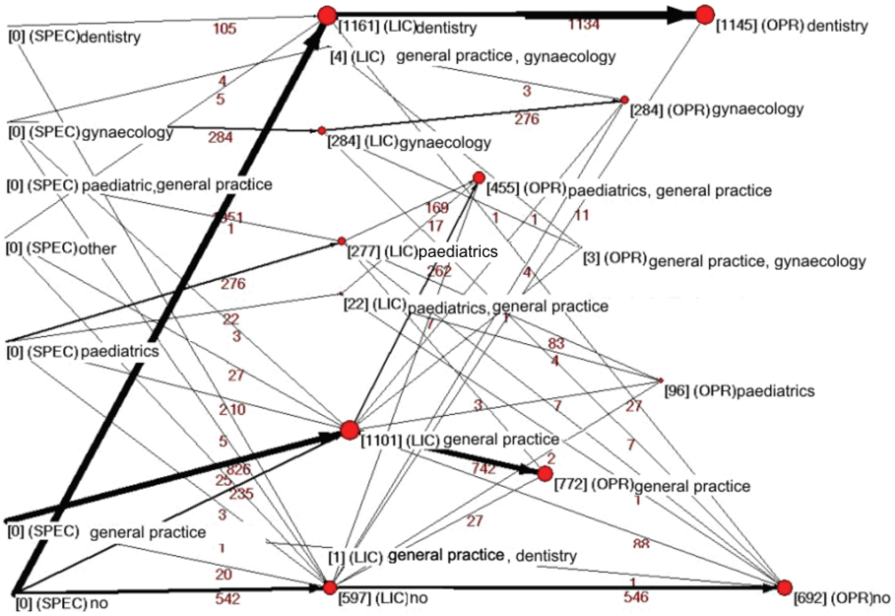


Figure 9: Qualifications of physicians for job and their patients.

Monitoring process B3 is aimed at the monitoring of physicians and dentists' qualification for the job they actually perform. The main performance indicator is the physician's specialization degree, granted by the Slovenian Medical Chamber, which must be verified every seven years. This specialization degree is a prerequisite for getting a license for employment in an area of medicine.

To monitor the suitability of physicians for the job they perform we have used a social network visualization technique (Figure 9) available in the social analysis program Pajek (Batagelj and Mrvar, 2006). The monitoring of physicians' suitability is achieved by the monitoring of three variables: specialization (*SPEC*), license (*LIC*), and the type of patients that the physician is in charge of, categorized by patient type (*OPR*). This analysis is motivated by the observation that physicians with some specialization may get different licenses, and while some licenses assume that the physician will only deal with patients of the corresponding patient category, in reality she may be in charge of different types of patients. For example, a pediatrician may provide health services to adult patients, although she has a specialization in pediatrics and a license in pediatrics.

The Pajek diagram (Figure 9) clearly shows the typical (thick lines – a high number of physicians) and atypical (thin lines – a low number of physicians) cases, which enable abnormality detection and further individual analysis of discovered anomalies. See for example the outlier indicated by the thin line connecting node [1101] (LIC) general practice and node [284] (OPR) gynecology. That line indicates one physician with a license of general practice who works as a gynecologist. Such outliers are worth investigating, as they could indicate severe irregularities in the PHCN, or alternatively indicate errors in the gathered data.

5.3 Visits

The module *Visits* addresses the frequency of patient visits to HCPs. Thus, MP C1 shows an average visit rate (*VR*) of listed patients from a municipality. *VR* is the ratio between the number of visits of listed patients (v_c) from municipality *c* and the number of listed patients (p_c) in that municipality:

$$VR = \frac{v_c}{p_c} \tag{5}$$

Monitoring process C1 is further improved by MP C2, which is focused on the ratio between the actual visit rate and the expected visit rate of the same listed patients over the same period. The expected visit rate for each age-gender group is calculated by the average number of visits. Hence, the patients are grouped by age and gender, and each group is weighted according to the average number of visits to HCP.

The output data of MP C2 is presented by two data visualization methods, which are used in order to simplify the detection of unusual visit rates of some groups of patients. The first visualization is a 3D bar chart that shows the ratio between actual visit rates to HCPs and expected rates for patients grouped by age and municipalities (Figure 10). In the chart, the vertical axis shows the ratio between actual visit rates and the expected rates. The chart clearly reveals unusually high visit rates of population under 20 to the general practice in some municipalities, especially in Luče and Nazarje, compared to other municipalities (not shown in this figure, due to space constraints). This anomaly may be explained by the lack of pediatricians in these municipalities, which is compensated by an increased number of visits to general practitioners.

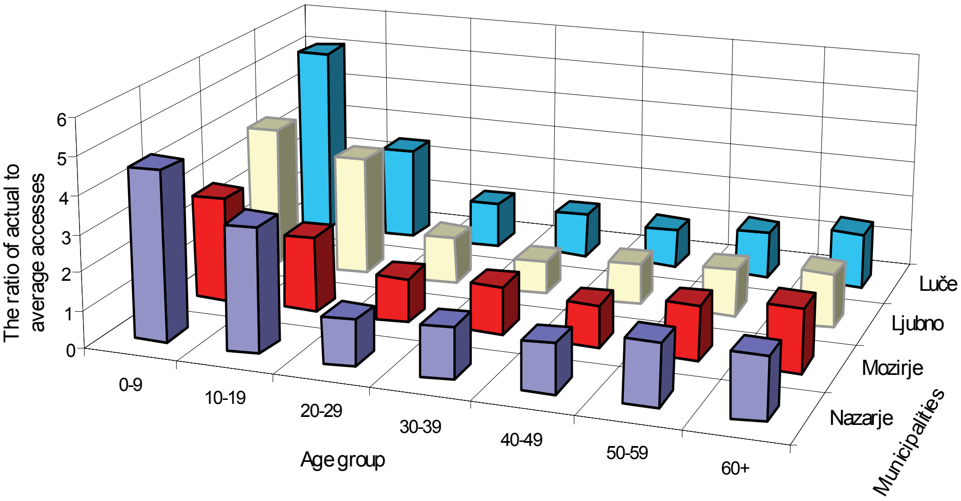
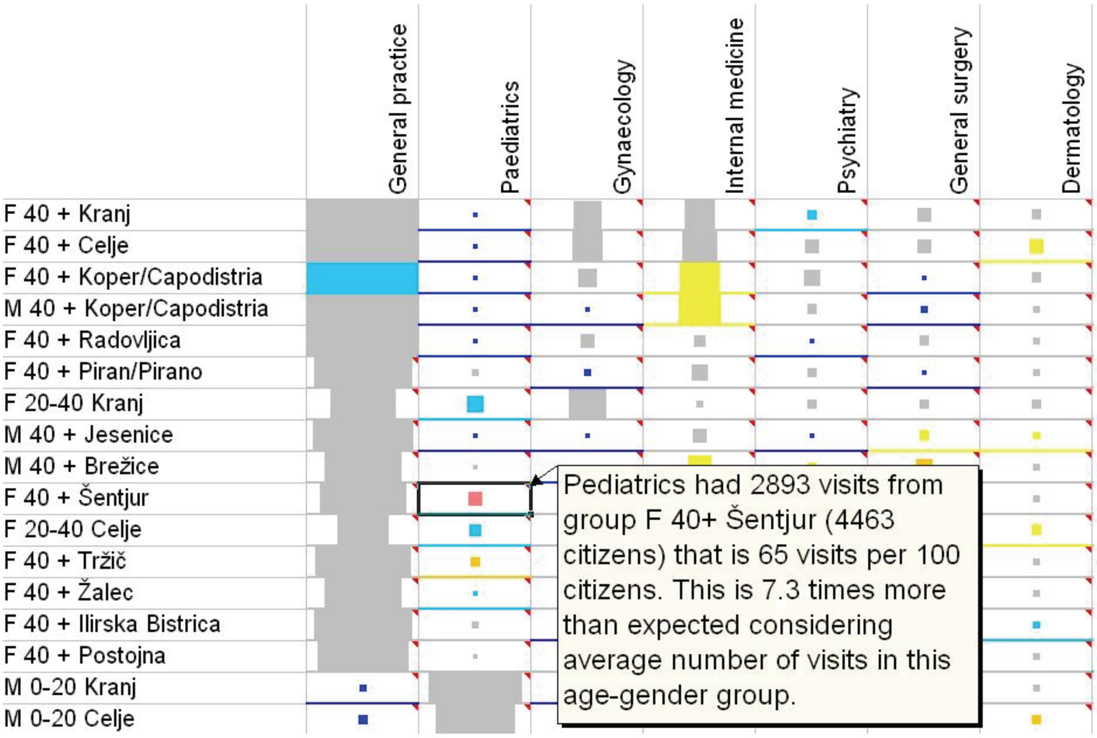


Figure 10: The ratio between actual and average visits of patients to general practice HCPs (year 2003).

The second data presentation method used in MP C2 shows associations between the visit rates of age-gender-grouped patients from municipalities and patients grouped by medical specialists. In this case, we investigate a limited number of association rules with a predefined structure. A clear visualization of all these rules can be achieved by a matrix of rectangles where the rows present age-gender-grouped patients from municipalities, specializations are presented by the columns and associations are presented by rectangles (Figure 11). The size of each rectangle is proportional to the parameter *Support* (the number of visits of grouped population to physicians that perform some specialization, see Section 3.2.1), and the color of the rectangle depends on the parameter *Lift* (the ratio of actual number of visits and the expected number of visits, Section 3.2.1). The red rectangles represent strong relations (with *Lift* higher than 4), and blue ones represent associations where the *Lift* is smaller than one.



Pediatrics had 2893 visits from group F 40+ Šentjur (4463 citizens) that is 65 visits per 100 citizens. This is 7.3 times more than expected considering average number of visits in this age-gender group.

Figure 11: Association rules presented by a matrix of rectangles.

A detailed explanation of each relation can be provided in the message box obtained by moving the screen pointer on the selected rectangle (Figure 11). This kind of visualization can display a large number of relations on a single screen (in our case, more than 10,000 relations). With this method some interesting relations can be found at a glance and be instantly explained.

In general, this visualization technique is appropriate for the identification of outliers. For instance, an increased visit rate of patients to HCPs can indicate unhealthy living environments or unhealthy habits of some population groups. Consequently, in addition to PHCN planning, the visit rates are also important for monitoring the health level of the population.

5.4 Accessibility

The module *Accessibility* includes two MPs aimed at monitoring the physical accessibility of patients to health care providers. MP D1 measures the *Accessibility for Patients (AP)*, which shows the proportion of the population that needs excessive time to get to the nearest HCP than is the maximal acceptable travel time, considering the length and the category of roads. Thus, the maximal acceptable travel time for the patients to the nearest HCP have to be chosen carefully. In our case, we set the maximal acceptable travel time statistically as one standard deviation above the national average. Generally, this time must be prescribed by the responsible authority. The *AP* for 2004 is presented by the table (Table 2) and the map (Figure 12).

Municipality	General Practice	Pediatrics	Gynecology
Celje	18.4%	3.8%	6.1%
Gornji Grad	50.7%	21.6%	93.2%
Kozje	66.3%	46.4%	100.0%
Laško	43.2%	30.1%	24.4%
Ljubno	19.5%	3.8%	3.2%
Luče	29.6%	23.1%	100.0%
Mozirje	34.0%	34.4%	100.0%

Table 2: Accessibility for Patients (*AP*): the proportion of population in a municipality that needs more time to get to the nearest HCP than prescribed.

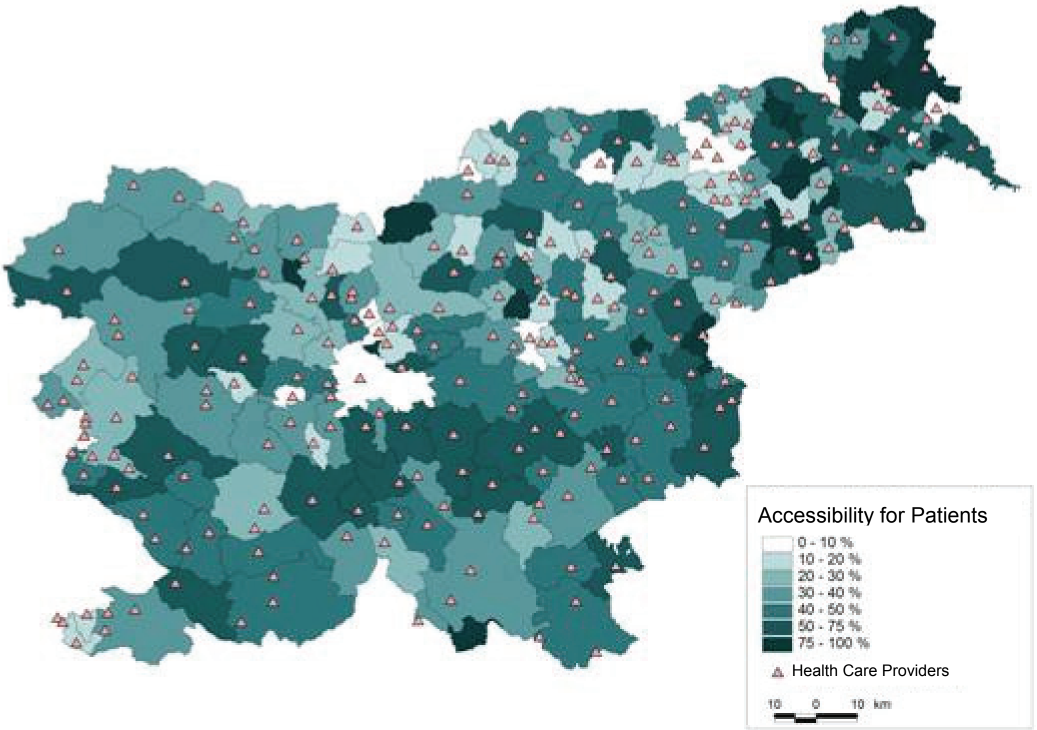


Figure 12: Access of patients to the nearest general practice HCPs (*AP*).

The next MP D2 shows the color-coded overview of *Full Accessibility* for Patients (*FAP*). It assesses the proportion of the population that needs more time than prescribed to get to the nearest HCP that is open 24 hours. The access time computation takes into account the road length and the road category. The difference between *FAP* and *AP* is that the first one considers only the HCPs working 24 hours, and the second one considers all the public health care providers. Because the accessibility of HCPs is important for patients, the MPs D1 and D2 are used as formal indicators, which affect the decisions related to health care resource allocation.

5.5 Unlisted patients

The module *Unlisted Patients* monitors the proportion of unlisted habitants in municipalities. The unlisted patients are those without their own physicians. The proportions of unlisted population, provided by MP E1, are calculated as the ratio between the number of unlisted patients and the entire population in a municipality. This MP provides important information about the patients in a municipality, thus it is used as a formal indicator.

5.6 Primary Health Care Network Availability

This module is focused on the assessment of the availability of the PHCN for listed patients. This assessment shows how the HCP capacity is adapted for the patient needs. The providers in the PHCN with insufficient capacity are overloaded, and their patients are health care deprived. On the other hand, the providers with excess capacity increase the costs of the PHCN. Generally, the assessment of the availability is based on the ratio of the PHCN capacity available for patients and the demand for this capacity from the same patients.

Thus, the MP F1 provides *Network Availability for Visits* (*NAV*) that is aimed at the assessment of the PHCN availability considering the visits of the listed patients and the PHCN working time. The PHCN working time is the time when the physicians of HCPs included in the network are available for their patients. The *NAV* is proportional to the average PHCN working time that could be spent for the visit of patients from a certain area.

$$NAV = \frac{1}{v_c} \sum_i PAV_i v_{ci} \tag{6}$$

The simplest assessment of the network availability for the visits is based on the ratio of PHCN working time in a municipality to total patients' visits from the same municipality. This assessment does not take into account that many patients access HCPs in the neighboring or even more distant municipalities. Moreover, some municipalities do not have their own HCPs at all. Since the migration of patients into other municipalities is an important factor, we designed the formula that takes these migrations into account (Formula 6). Thus, *NAV* is the coefficient between the estimated PHCN working time spent for the visits of the patients from the municipality *(c)* and the total number of visits (v_c) from the patients living in the same municipality. The estimated PHCN working time is the sum of products between PAV_i and the number of patients' visits (v_{ci}) from the municipality *(c)* to the HCP (*i*) (Formula 6). Here, the *Providers' Availability for Visits* (PAV_i) (see MP H1) means the average working time of the *i*-th HCP available for the visit.

Because the *NAV* indicator does not consider that different patient treatments require different working time, this assessment is improved by MP F3, which gives the *Network Availability for Adjusted Visits* (*NAAV*). The *NAAV* indicator is proportional to the ratio between the PHCN adjusted working time spent for visits from the municipality (*c*) and adjusted visits (v_{ac}) from the same municipality (Formula 3).

$$NAAV = \frac{1}{v_{ac}} \sum_i PAAV_i v_{aci} \tag{7}$$

Here, the PHCN adjusted working time is calculated as the sum of products between providers' availability for adjusted visits (*PAAV*) (see MP H3) and adjusted visits (v_{aci}) from municipality (*c*) to provider (*i*) (Formula 7). The adjusted visits v_{ac} are proportional to the working time spent for overall visits of patients from municipality (*c*), and adjusted visits v_{aci} are proportional to the working time spent for visits from municipality (*c*) to the provider (*i*). The visits are adjusted, because different treatments of patients spend different working time. Therefore, they are grouped by health treatments, and each group is weighted in accordance with the expected physician's working time spent for visits.

The next MPs F2 and F4 are focused on the listed patients. The MP F2 indicator provides the *Network Availability for Listed patients* (*NAL)* aimed at measuring the availability of the PHCN for listed patients from a certain area. The *NAL* indicator is proportional to the average PHCN working time that is available for a listed patient in a certain period (usually one year). The simplest assessment of the PHCN availability for the listed patients from a municipality is based on the ratio of the PHCN working time of HCPs in a municipality to total number of listed patients from the municipality. Because this assessment does not take into account the migration of patients into other municipalities, we designed a new assessment of the PHCN availability for listed patients. In this case, the *NAL* is defined as the ratio of the PHCN working time available for listed patients from the municipality and the number of the same listed patients (p_c) (Formula 8). The available PHCN working time is defined as the sum of the products between providers' availability for patients (*PAP*) (see MP H2) and patients (p_{ci}) from municipality (*c*) listed for the given provider (*i*) (Formula 8).

$$NAL = \frac{1}{p_c} \sum_i PAP_i p_{ci} \tag{8}$$

Similarly to the *NAV* indicator, the *NAL* indicator does not consider that different age-gender groups of listed patients have a different workload PHCN. Hence, in MP F4, the *NAL* indicator is upgraded so that each age-gender group of listed patient is adjusted in accordance with the expected physician's working time spent for treatments. The output of this process is called *Network Availability for Adjusted Listed patients* (*NAAL*). The *NAAL* is proportional to the ratio between available working time of PHCN for adjusted listed patients from a municipality (*c*) and the needed working time for adjusted listed patients (p_{ac}) from the same municipality (Formula 9). The available working time of PHCN for adjusted listed patients is the sum of the products between providers' availability for adjusted patients (*PAAP*) (see MP H4) and needed working time for adjusted patients (p_{aci}) from that municipality listed on HCP (*i*).

$$NAAL = \frac{1}{p_{ac}} \sum_{i} PAAP_i \, p_{aci} \qquad (9)$$

The MPs based on patients' visits, such as F1, F3, H1, and H3, show the availability of the PHCN in the past. In contrast, the MPs based on the number of listed patients, such as F2, F4, H2, and H4, provide an anticipated availability of the PHCN. This anticipation considers the characteristics of listed patients such as gender and age. In general, the assessment of workload for the past shows the anomalies in the PHCN that have already happened; on the other side, the anticipated assessment of workload provides the information that could prevent future anomalies in the PHCN, but they are usually less accurate.

The MPs F3 and F4 need the coefficients that weight age-gender grouped patients and their visits grouped by treatments in accordance with the expected workload for physicians. These coefficients have not been confirmed yet by the Ministry for Health. Thus, in our model the MPs F3 and F4 are only monitored, while the formal MPs F1 and F2 are propagated to higher levels.

The MPs in this module show the availability of PHCN only for listed patients and not for the whole population. Therefore, the anomalies in the PHCN that could derive from a higher number of unlisted patients is provided by the MP E1, which is aimed at the ratio of unlisted to all patients in municipalities.

The output data from MPs in this module are usually presented by tables and charts based on OLAP techniques, and maps based on GIS techniques (MediMap, 2004; Pur, et al., 2005a).

5.7 Dispersion

This module is aimed at monitoring the dispersion of locations where physicians work. Depending on the requirements, a physician may work in more than one location, but this dispersion usually means additional workload for physicians and their lower availability for patients at one of the locations. The dispersion is provided by MP G1 and means the average number of locations where physicians in the municipalities work. It can be presented by a map such as the one shown in Figure 13, which shows the physicians' dispersion in year 2004 for gynecology. The darker municipality has a higher average number of locations where physicians from this municipality work.

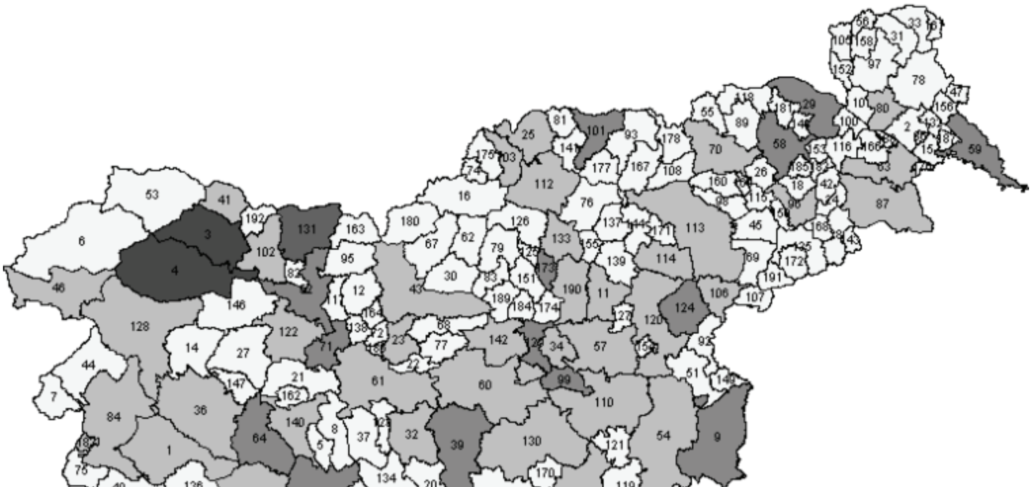


Figure 13: Average dispersion of working locations of gynecologists in Slovenia (year 2004).

5.8 Health Care Providers' Availability

The module *HCP Availability* is focused on the assessment of the HCPs and physicians' availability for their patients. This assessment shows how the capacity of HCPs is adapted for the patients' needs. The HCPs with insufficient capacity are overloaded, and their patients are health care deprived. On the contrary, the excess capacity increases the costs of providers. Generally, this assessment is based on the ratio between the capacity of HCP and the needs of the patients listed on this provider. The main difference between the monitoring processes in this module and processes in module *PHCN Availability* is that the former processes are aimed at the assessment of the availability of HCPs for patients listed on this provider, and processes in the latter module are aimed at the assessment of the availability of the PHCN for patients from a certain region.

MP H1 is aimed at the assessment of the HCP *Availability for Visits* (*PAV*) considering the patient's visits to HCP and HCP working time. The HCP working time is the time when the provider's physicians are available for patients. The *PAV* is proportional to the average HCP working time that can be spent for a patient's visit. Thus, the *PAV* is based on the ratio of the HCP working time (t_i) to number of visits (v_i) to this provider (i) in the same period (Formula 10).

$$PAV = \frac{t_i}{v_i} \tag{10}$$

The similar MP H3 gives *Provider Availability for Adjusted Visits* (*PAAV*) aimed at the availability of a HCP for its patients considering the visits adjusted in accordance with the expected physician's working time needed for treatments. The *PAAV* is defined as the ratio of the available working time (t_i) of the HCP (i) for their patients in a certain period to adjusted visits (v_{ai}) in the same period (Formula 11). The adjusted visits (v_{ai}) are proportional to the anticipated working time spent for all visits to a provider (i). Therefore the visits are grouped in accordance with the health treatment where each group is weighted considering the expected working time spent for the treatment.

$$PAAV = \frac{t_i}{v_{ai}} \tag{11}$$

The next MPs H2 and H4 are focused on listed patients. Thus, MP H2 provides the assessment of the *Providers Availability for Patients* (*PAP*). The *PAP* is proportional to the average HCP working time that can be spent for a listed patient in a certain period (usually one year). It is defined as the ratio of the HCP (i) working time (t_i) in a certain period to number of patients (p_i) listed to the provider (i) in the same period.

$$PAP = \frac{t_i}{p_i} \tag{12}$$

PAP does not consider that different age-gender groups of listed patients require different HCP workloads. Hence, this assessment is improved by the next MP H4 where listed patients are adjusted in accordance with the expected physician's working time spent for their treatments. We call this assessment *Providers' Availability for Adjusted Patients* (*PAAP*). The *PAAP* is proportional to the ratio between the working time (t_i) of HCP (i) and expected working time needed for the listed patients (p_{ai}). The estimation of expected working time is based on the age-gender structure of the listed patients. Here the listed patients are age-gender grouped, and each group is weighted according to the expected physician's working time spent for their treatments.

$$PAAP = \frac{t_i}{p_{ai}} \tag{13}$$

The results provided by the MPs in this module can be presented by different methods such as tables based on OLAP technologies and map-based visualizations. Again, for brevity, the results are excluded from this paper.

The monitoring of physicians' and HCP's availability for patients *PAV* and *PAAV*, which are based on patient visits, may induce unwanted side effects. For example, to improve their *PAV* and *PAAV*, physicians or HCPs may arrange more visits to their patients than necessary in order to show their lower availability for patients and higher workload. This manipulation can be prevented by indicators (*PAP* and *PAAP*) based on the number of listed patients.

5.9 Health Activities

The module *Activities* provides the basic data about activities done by HCPs. The activities that we analyze are the patient's visits to HCPs. The data about these visits, such as time, HCPs, and diagnosis, are provided by MP H1. MP H2 provides the visits adjusted according to the expected time for health treatment. In H2, the visits are grouped in accordance with the expected time. Each group is weighted considering the physician's working time spent for the treatment. Thus, the adjusted visit is proportional to the physician's time spent for health treatment. The output data from these MPs can be combined with other data or presented by different methods.

5.10 Population

The module *Population* is focused on data about the population serviced by the HCPs included in the PHCN. The first MP J1 provides different data about the population aggregated across municipalities where they live. Most of this data was obtained by the Statistical Office of the Republic of Slovenia. The second MP J2 provides data about listed patients such as gender, age, and the municipalities where they live. Most of this data was obtained from the Institute of Public Health of the Republic of Slovenia. This module also includes MP J3, which provides the expected working time of physicians spent for a listed patient. The listed patients are grouped according to their age and gender, and each group is weighted in accordance with the expected working time spent for all treatments in a certain period. Usually this period is one year.

5.11 Health Resources

This module provides the basic data about physicians and HCPs. This is done by the MP I1, which provides the different data about physicians such as age, gender, HCP where they work, FTE, and education. The main sources of this data are the National Institute of Public Health, Health Care Institute Celje, the Slovenian Social Security database, and the Slovenian Medical Chamber.

6 Original Scientific Papers

6.1 Data Mining for Decision Support: An Application in Public Health Care

The title of the article in this chapter is *Data Mining for Decision Support: An Application in Public Health Care* (Pur, et al., 2005a). The article describes the results of the MediMap (2004) project, where we developed methods and tools for management improvement of the Health Care Network (HCN) in the Celje region.

The article describes the DM techniques used for the clustering of Community Health Centers, and the creation of the assessment models for the identification of anomalies related to health care capacities and patients' needs in the municipalities.

The article was presented at the 18th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, IEA/AIE 2005, Bari, Italy, June 22–24, 2005, and was published in the Springer series Lecture Notes in Computer Science.

Data Mining for Decision Support: An Application in Public Health Care

Aleksander Pur[1], Marko Bohanec[2,5], Bojan Cestnik[6,2],
Nada Lavrač[2,3], Marko Debeljak[2], and Tadeja Kopač[4]

[1] Ministry of the Interior, Štefanova 2, SI-1000 Ljubljana, Slovenia
aleksander.pur@policija.si
[2] Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia
{marko.bohanec, nada.lavrac, marko.debeljak}@ijs.si
[3] Nova Gorica Polytechnic, Nova Gorica, Slovenia
[4] Public Health Institute, Celje, Slovenia
[5] University of Ljubljana, Faculty of Administration, Ljubljana, Slovenia
[6] Temida, d.o.o. Ljubljana, Slovenia

Abstract. We propose a selection of knowledge technologies to support decisions of the management of public health care in Slovenia, and present a specific application in one region (Celje). First, we exploit data mining and statistical techniques to analyse databases that are regularly collected for the national Institute of Public Health. Next, we study organizational aspects of public health resources in the Celje region with the objective to identify the areas that are atypical in terms of availability and accessibility of the public health services for the population. The most important step is the detection of outliers and the analysis of the causes for availability and accessibility deviations. The results can be used for high-level health-care planning and decision-making.

Keywords: Data Mining, Decision Support, Knowledge Discovery, Knowledge Management, Applications to Health Care.

1 Introduction

Effective medical prevention and good access to health care resources are important factors that affect citizens' welfare and quality of life. As such, these are important factors in strategic planning at the national level, as well as in planning at the regional and local community level. Large quantities of data collected by medical institutions and governmental public health institutions can serve as a valuable source of evidence that should be taken into account when making decisions about priorities to be included in strategic plans.

The organization of public health care in Slovenia is hierarchical: the national Institute of Public Health (IPH) coordinates the activities of a network of regional Public Health Institutes (PHIs), whose functions are: monitoring public health, organizing the public health activities, and proposing and implementing actions for maintaining and improving public health. PHIs themselves coordinate a regional network of hospitals, clinics, individual health professionals and other health care resources. The

M. Ali and F. Esposito (Eds.): IEA/AIE 2005, LNAI 3533, pp. 459 – 469, 2005.

system of public health is thus organized at three levels: strategic (the Ministry of Health and the national IPH), managerial (regional PHIs) and operational (local hospitals, clinics, individual health professionals and other health care resources).

The network of regional PHIs, coordinated by the national IPH, collects large amounts of data, which require appropriate *knowledge management* [1]. Knowledge management is recognized as the main paradigm for successful management of networked organizations, aimed at supporting business intelligence [2] – a broad category of applications and technologies for gathering, storing, analysing, and providing access to data to help organizations make better decisions. In addition to the technological solutions, it needs to address organizational, economic, legislative, psychological and cultural issues [3].

Knowledge management can be supported by the use of knowledge technologies, in particular by *data mining* and *decision support* [4], which are in the focus of the work described in this paper. Data mining and decision support have a large potential for knowledge management in networked organizations, and have already proved to be successful in numerous applications. Data mining is typically applied to knowledge discovery in large and complex databases and has been extensively used in industrial and business problem solving, while its use in health care is still rare. In such a knowledge intensive domain, neither data gathering nor data analysis can be successful without using knowledge about both the problem domain and the data analysis process, which indicates the usefulness of integrating data mining with decision support techniques to promote the construction of effective decision criteria and decision models supporting decision making and planning in public health care.

This paper describes an application of data mining and decision support in public health care, which was carried out in Slovenia within a project called MediMap. Section 2 briefly overviews the two research areas, data mining and decision support, and proposes their integration to better solve data analysis and decision support problems. Section 3 presents the specific application of these techniques, which was developed for the Public Health Institute of the Celje region.

2 Data Mining and Decision Support in Knowledge Management

Data mining [5,4] is concerned with finding interesting patterns in data. Data mining includes predictive data mining algorithms, which result in models that can be used for prediction and classification, and descriptive data mining algorithms for finding interesting patterns in the data, like associations, clusters and subgroups.

Decision support [6,4] is concerned with helping decision makers solve problems and make decisions. Decision support provides a variety of data analysis, preference modelling, simulation, visualization and interactive techniques, and tools such as decision support systems, multiple-criteria modelling, group decision support and mediation systems, expert systems, databases and data warehouses. Decision support systems incorporate both data and models.

Data mining and decision support can be integrated to better solve data analysis and decision support problems. In *knowledge management* [1], such integration is interesting for several reasons. For example, in data mining it is often unclear which algorithm is best suited for the problem. Here we require some decision support for data mining. Another example is when there is a lack of data for the analysis. To ensure that appropriate data is recorded when the collection process begins it is useful to first build a decision model and use it as a basis for defining the attributes that will describe the data. These two examples show that data mining and decision support can complement each other, to achieve better results. Different aspects of data mining and decision support integration have been investigated in [4].

In MediMap, we mainly used descriptive data mining methods, and combined them with visualization and multiple-criteria techniques, as shown in the next section.

3 Data Mining and Decision Support: Health-Care Application

The main goal of the project MediMap was to establish a knowledge repository for supporting decisions in the field of planning the development of community health care centres (CHC) for a regional PHI Celje. We approached this goal in two phases: first, we analysed the available data with data mining techniques, and then, we used the acquired understanding of the data and the domain as leverage for a more elaborate study of the field with decision support techniques.

In the first phase, using data mining techniques, we focused on the problem of directing patients from the primary CHCs to the specialists. The main assumption was that similar CHCs should have comparable directing rates. For the similarity measure we took patients' age and social categories, as well as organization and employment structure of the CHCs. The results revealed that the deliberate aggregation of data, although justified for the primary purpose of data gathering, probably hid most of the interesting patterns that could be exposed in the data mining phase. Consequently, the need for additional data gathered from CHCs was forwarded to the national IPH. This data could be obtained at almost no additional costs, since it is already collected by CHCs, but aggregated too early in the data acquisition and reporting process. At the same time, we gained a substantial insight into the domain, which served as reinforcement for the further studies.

In the second phase we studied organizational aspects of public health resources in the Celje region. The goal was to identify the areas that are atypical in terms of availability and accessibility of the public health services for the population, which could provide valuable information to support decisions related to planning the future development of public health services. For the estimation of parameters from data, we used the same database as in the first phase. Additionally, we derived a model for estimating the availability and accessibility that incorporates several innovative criteria. Moreover, we gathered additional geographic information from several other data-sources, like statistical data for the population of a given area and distance measures between cities.

The most important step of the second phase was the detection of outliers and the analysis of the causes for different availability and accessibility figures. The result of the described process is summarized in Fig. 7, which can be used as a high-level information fusion tool for planning the requirements for the employees for health care services in the Celje region.

3.1 Analysis of Health Care Centres Data with Data Mining

First, we have tried to set up appropriate models and tools to support decisions concerning regional health care in the Celje region, which could later serve as a model for other regional PHIs. The requirements, formulated by the PHI Celje, fall into three problem areas: (1) health care organization (the PHI network, health care human resource distribution), (2) accessibility of health care services to the citizens, and (3) the network of health care providers. These requirements were made operational as five problem tasks:

- analysis of the public health providers network,
- analysis of public health human resources,
- analysis of public health providers workload,
- management and optimisation of the public health providers network, and
- simulation and prediction of the performance of the public health providers and human resource network.

The dataset for the analysis consisted of three databases: (1) the health care providers database, (2) the out-patient health care statistics database (patients' visits to general practitioners and specialists, diseases, human resources and availability), and (3) the medical status database.

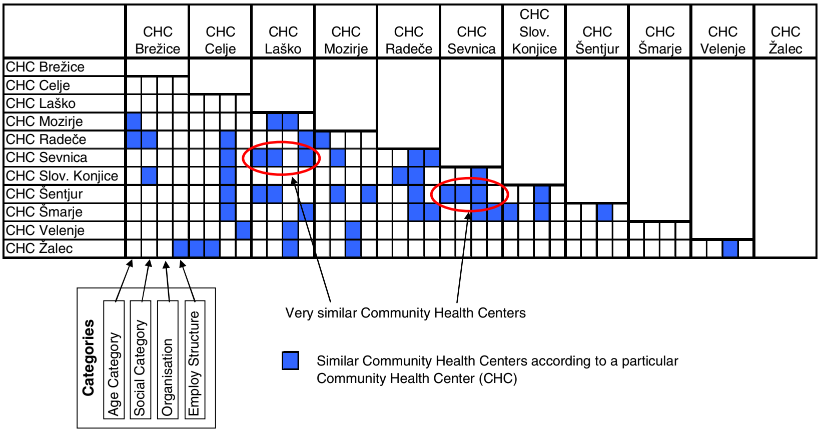


Fig. 1. The similarity matrix of community health centres (CHCs) in Celje

To model the processes of a particular CHC (the patient flow), data describing the directing of patients to other CHCs or specialists were used. Our intention was twofold: to detect the similarities between CHCs, and to detect the atypical CHCs. Similarities between CHCs were analysed according to four different categories: patient's age categories, patient's social categories, the organization of the community health centre, and employment structure of the community health centre. For each category, similarity groups were constructed using four different clustering methods: agglomerative classification [7], principal component analysis [7], the Kolmogorov-Smirnov test [8], as well as the quantile range test and polar ordination [9]. Averages over four clustering methods per category were used to detect the similarities between the CHCs of the Celje region (Fig. 1).

These results were evaluated by domain experts from PHI Celje. In several cases the results confirmed already known similarities, while the experts could not find any reasonable explanations for new knowledge described in the similarity matrix, as the data describing was too coarse (aggregated).

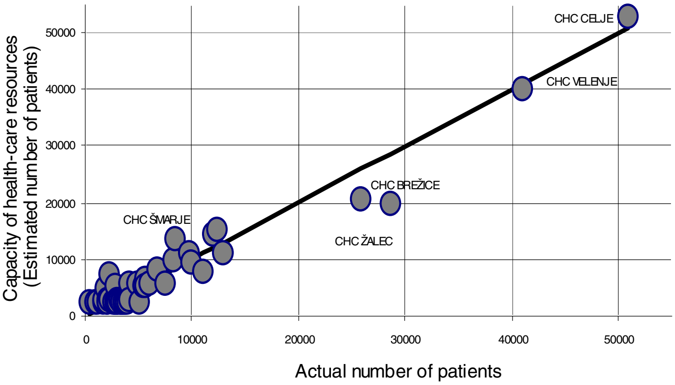


Fig. 2. Detecting atypical health care resources in Celje

The analysis of the typicality of CHCs was based on the comparison of the estimated number of patients that can be handled by a CHC (its capacity estimated by the number of employed staff), and the actual number of patients handled by the CHC. The outcome, which is shown in Fig. 2, was well appreciated by the experts. The figure presents some atypical CHCs, deviating from the diagonal line, such as CHC Brežice and Žalec, which have an insufficient number of staff compared to the number of actual patients.

3.2 Availability and Accessibility of Public Health Care Resources

The goal of this analysis was to detect the local communities that are underserved concerning general practice health services – this means that that the population in these areas have available less than a generally accepted level of services. We evaluated 34 local communities in the Celje region. The evaluation is based on the ratio of the capacity of health care services available to patients from the community and the demand for these services from the population of the same area. In our case, the *capacity* ability of health care services is defined as available time of health care services for patients in that community, and *demand* means the number of accesses to health care services from patient from the community. Therefore, our main criterion for the evaluation of health care system for patients in community c is actually the average time available in health services per access of a patient from this community. We call this criterion *AHSP* (Availability of Health Services for Patients):

$$AHSP = \frac{\sum t_i}{p_c} \tag{1}$$

Here, t_i denotes the total working time of health-care service i in community c, and p_c the number of accesses to health care services of patients from the community c.

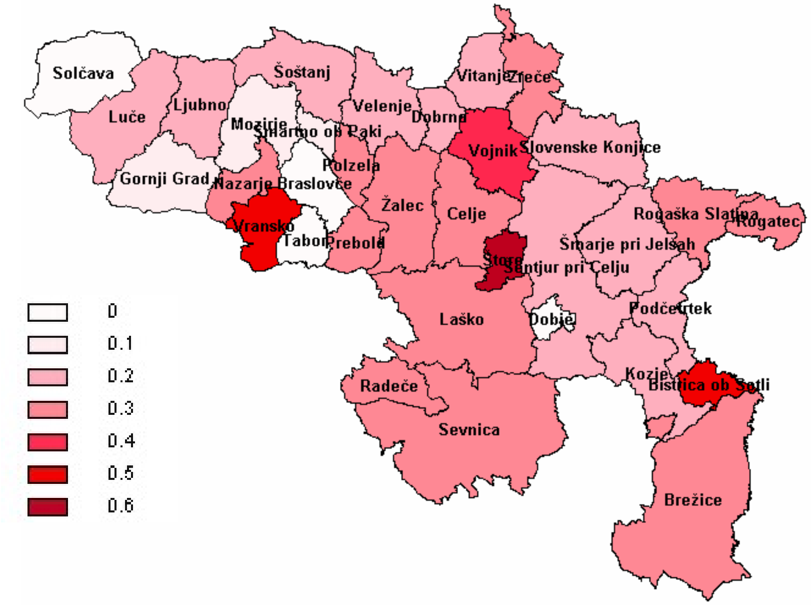


Fig. 3. Availability of health services (*AHSP*), measured in hours, in the Celje region in 2003

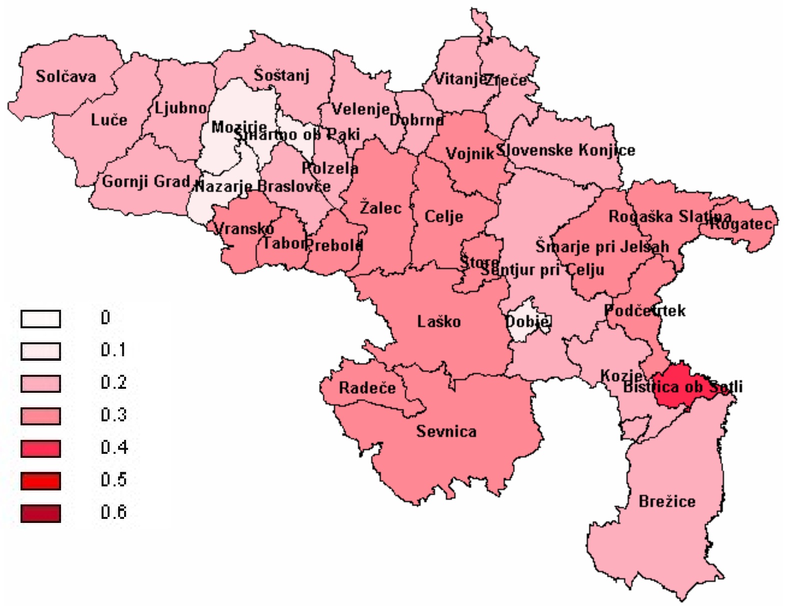


Fig. 4. Availability of health services in Celje in 2003, measured in hours, considering the migration of patients to neighbouring communities ($AHSP_m$)

AHSP does not take into account that many patients access health services in neighbouring communities. Moreover, some of communities do not have their own health care services at all. The migration of patients into neighbouring communities is considered in the criterion $AHSP_m$ as follows:

$$AHSP_m = \frac{1}{p_c}\sum_i a_i p_{ci} \tag{2}$$

Here, a_i is the *available time* of health care service *i* per person, defined as the ratio of total working time of health-care services and total number of visits, p_c is the total number of accesses of patient from the community *c*, and p_{ci} is the number of accesses of patients from the community *c* to health service *i*.

The evaluation of some communities in the Celje region using the criteria *AHSP* and $AHSP_m$ is shown in Fig. 3 and Fig. 4, respectively. The colour of communities depends on the availability of health services for patients: the darker the colour, the higher the health care availability in the community (measured in hours). The main difference between the evaluations is noticeable in communities without their own health care services, like Braslovče, Tabor, Dobje and Solčava. If the migration of patients in neighbouring communities is not considered, then it looks like that the inhabitants of these communities are without health care (Fig. 3). Thus, $AHSP_m$ (Fig. 4) provides a more realistic evaluation criterion. Such a geographical representation of the results has been extremely well accepted by the health-care experts.

For even a clearer picture about the availability of health-care services and for the purpose of its visualization (Fig. 5), we introduce two additional criteria. The criterion *AHS* (Availability of Health Services) is defined as the availability of health care services for the population from community c. More precisely, *AHS* is defined as the available time of health care services per population g_c from the community *c*, considering the migration:

$$AHS = \frac{1}{g_c}\sum_i a_i p_{ci} \tag{3}$$

The next criterion, *RAHS* (Rate of Accesses to Health Services), defines the rate of accesses to health care services for population g_c from the community *c*:

$$RAHS = \frac{p_c}{g_c} \tag{4}$$

In this case, $AHSP_m$ is defined as the ratio between the availability of health services for population from community and the rate of visiting health service:

$$AHSP_m = \frac{AHS}{RAHS} \tag{5}$$

All these criteria give us some very interesting indicators about health conditions and health care in communities. They can be conveniently presented as shown in Fig. 5. Four measurements are actually shown in the chart: *RAHS* along the horizontal axis, *AHS* along the vertical axis, $AHSP_m$ as dot colour, and the population size (g_c) as dot diameter. Communities with average values of *RAHS* and *AHS* appear in the middle of the chart. The outliers represent more or less unusual communities regarding health care. Communities on the left side of the chart have lower rate of access to health services and the ones on the right side have higher accessing rate. On the bottom are located the communities with lower values of *AHS* and on the top with higher. The dark-coloured communities have higher values of $AHSP_m$ than the light-coloured ones.

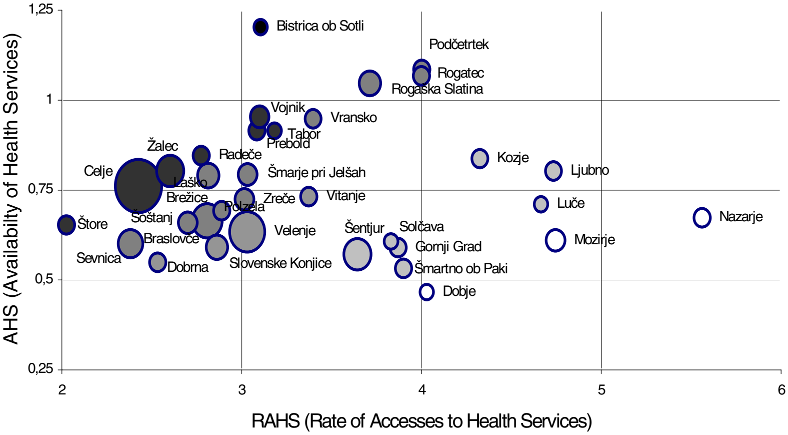


Fig. 5. Available time of health care services per population by community (2003)

Thus, Fig. 5 enables discovering implicit and interesting knowledge about health in communities. For example, the reason for high value of $AHSP_m$ in communities on the left side of the chart (e.g., Štore) could be the low rate of accesses to nearest health services, caused by inappropriate cure in these services. The reason for the low value of $AHSP_m$ in communities on the right side (Nazarje, Mozirje, Luče in Ljubno) is high rates of accesses to health services.

3.3 Decision Support for Planning Health Care Resources

Additional explanation of these rates can be provided by a chart as shown in Fig. 6. It shows the ratio of actual rate of accesses of health services and expected rate for age group of population in communities. This ratio is used in order to simplify detecting unusual rate of accesses to health services. The expected rate of accesses to health services is the average rate of population in age group. For example, the access to health services from population between 60 and 69 are almost five times as frequent as these between 20 and 29. The age group of population from communities is measured along the horizontal axis. Thus, the chart shows that the characteristic for these communities is unusual high rates of accesses to the health services of population under 20. Therefore we could presume that the main reason for the high value of $AHSP_m$ in these communities is the absence of paediatric services.

Further view on the disparity of health care in communities (Fig. 5) is provided in Fig. 7. There, the evaluation of health services is based on the ratio between the health-care capacity and demand. In our case the demand means the number of accesses to health services, and is measured along the horizontal axis. Capacity is proportional to the working time of health services, and is measured along vertical axis. Some of health services are denoted with identification number and community. Regression line in the chart represents the expected working time of health services, with respect to the number of accesses. The working time of the health services under the regression line, like Nazarje and Mozirje, is too short, and of these above the regression line is too long. Thus, this chart can serve for supporting decisions in planning the capacity and working time of health care services.

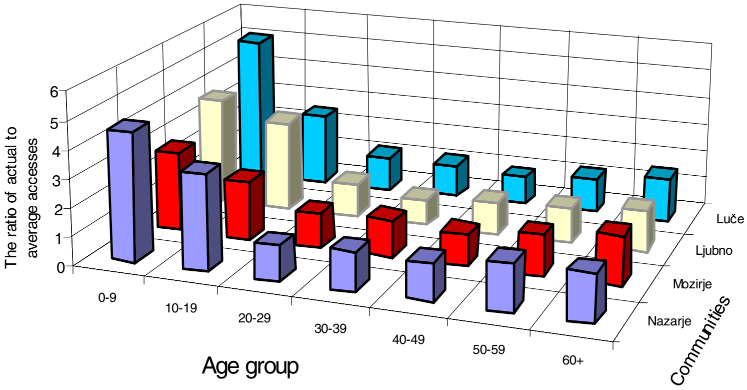


Fig. 6. The ratio between actual and average accesses (2003)

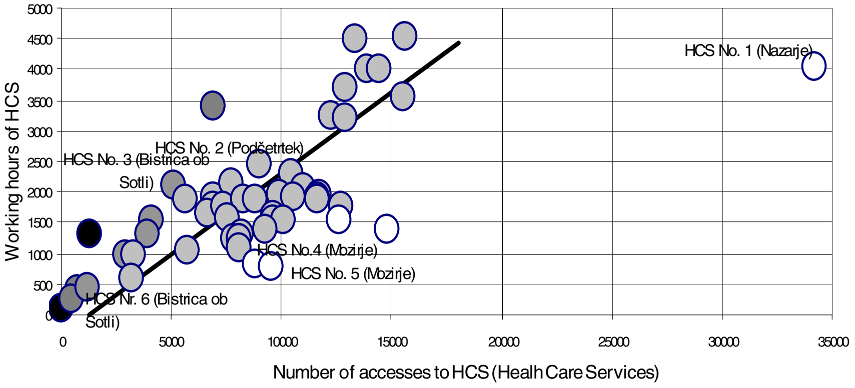


Fig. 7. The evaluation of health services: the ratio of health-care capacity and demand (2003)

4 Conclusion

Improved data mining and decision support methods lead to better performance in problem solving. More importantly, integrated data mining and decision support methods may further improve the performance of developed solutions and tackle new types of problems that have not been addressed before. A real-life application of this approach in public health care was shown in this paper.

In the MediMap project we have developed methods and tools that can help regional PHIs and national IPH to perform their tasks more effectively. Tools and methods were developed for the reference case of IPH Celje and tested on selected problems related to health care organization, accessibility of health care services to the citizens, and the health care providers work.

In the first part of the project, statistical and data mining methods were used in order to get acquainted with the problem domain and data sources. In the second part, we implemented decision support methods to the problem of planning the development of public health services. The main achievement was the creation of the model of availability and accessibility of the health services to the population of a given area. With the model it was possible to identify the regions that differ from average and to consequently explain the causes for such situations, providing many benefits for the health-care planning process.

In addition, the national IPH will use the results to identify missing data that should be included in the improved protocol of public health data gathering at the national level, as the study indicates that additional – more detailed, but relatively easy to obtain – data from the community health centres is needed. This finding is valuable for the IPH, which defines the national data model and prescribes data-gathering rules and procedures.

In further work, we will extend this analysis to other regions of Slovenia. We will focus on the development of decision support tools with the automatic modelling of health care providers using data mining. We wish to implement the developed methodology so that it can be regularly used for decision support in organisations reponsible for the health-care network: the Ministry of Health, the IPH, and PHIs.

Acknowledgements

We gratefully acknowledge the financial support of the Public Health Institute Celje, the Slovenian Ministry of Education, Science and Sport and the 6FP integrated project ECOLEAD (European Collaborative Networked Organizations Leadership Initiative). We also express our thanks to other members of the MediMap project team, in particular to Tanja Urbančič and Mitja Jermol, who have also contributed to the results described in this paper.

References

1. Smith RG, Farquhar A.: The Road Ahead for Knowledge Management: An AI Perspective. AI Magazine, Vol. 21, No. 4, 17–40 (2000)
2. Biere M.: Business Intelligence for the Enterprise. Prentice Hall PTR (2003)
3. McKenzie J, van Winkelen C.: Exploring E-collaboration Space. Henley Knowledge Management Forum (2001)
4. Mladenić D, Lavrač N, Bohanec M, Moyle S. (editors): Data Mining and Decision Support: Integration and Collaboration. Kluwer (2003)

5. Han J, Kamber M.: Data Mining: Concepts and Techniques. Morgan Kaufman (2001)
6. Mallach EG.: Decision Support and Data Warehouse Systems. McGraw-Hill (2000)
7. Legendre P, Legendre L.: Numerical Ecology. 317–341. Elsevier (1998)
8. Zar JH.: Bistatistical Analysis, 478-481. Prentice Hall (1999)
9. Ludwig JA, Reynolds JF.: Statistical ecology: A primer of methods and computing. Wiley Press, 337 (1988)

6.2 Monitoring Human Resources of a Public Health care System through Intelligent Data Analysis and Visualization

This article, describes a model for the monitoring and assessment of human resources in the Slovenian health care network (HCN) (Pur, et al., 2007a). The model is developed in accordance with the proposed methodology. It is composed of the following modules: human resources, age – workload, qualification, age, workload, and dispersion. Each module is composed of the monitoring processes (MP) that reveal various aspects of the human resources in the HCN.

The article was presented at the 11th Conference on Artificial Intelligence in Medicine, AIME 2007, Amsterdam, The Netherlands, July 7–11, 2007, and was published in the Springer series Lecture Notes in Computer Science.

Monitoring Human Resources of a Public Health-Care System through Intelligent Data Analysis and Visualization

Aleksander Pur[1], Marko Bohanec[2], Bojan Cestnik[4,2], and Nada Lavrač[2,3]

[1] Ministry of Interior Affairs, Ljubljana,
[2] Jožef Stefan Institute, Ljubljana,
[3] University of Nova Gorica, Nova Gorica,
[4] Temida, d.o.o., Ljubljana,
Slovenia

1. Introduction

According to the World Health Report (World Health Organization, 2000), a health-care system (HCS) is a system composed of organizations, institutions and resources that are devoted to producing a health action. Human resources are one of the main parts of this system.

This paper is focused on the model for monitoring and planning of human resources in the Slovenian public HCS. The HCS of Slovenia is divided into the primary, secondary and tertiary health-care levels. The primary health-care (PHC) is the patients' first entry point into the HCS. It is composed of four sub-systems: general practice, gynaecology paediatrics and dentistry.

We have developed a model for monitoring the network of physicians at the PHC level, taking into the account the physicians' specializations, their geographic and work-time dispersion, time capacity constraints and their availability for patients. The motivation for this development came from the Ministry of Health of the Republic of Slovenia, who need a holistic overview of the PHC network in order to make short- and long-term management decisions and apply appropriate management actions, as well as evaluate PHC target achievements.

The first step was to form a data warehouse; a corresponding Entity Relationship Diagram data model was composed of unique database entries from the following existing sources:

- Slovenian Social Security databases: the data about health-care providers together with assigned patients per individual general practitioner, assigned patients with social security, and the data about health-care centres,
- the database of Slovenian physicians and dentists (Slovenian Medical Chamber),
- the database of the National Institute of Public Health containing data about Slovenian health centres, and
- the database of the Slovenian Statistics Bureau concerning the demographic and geographic distribution of citizens and communities in Slovenia.

The next steps involved the development of the model for monitoring the network of primary-care professionals based on the established data warehouse. The model was used

Source: Data Mining in Medical and Biological Research, Book edited by: Eugenia G. Giannopoulou,
ISBN 978-953-7619-30-5, pp. 320, December 2008, I-Tech, Vienna, Austria

on real HCS data for the year 2006, which forms a part of a larger model for monitoring the responsiveness of the HCS for population, developed for the Ministry of Health of the Republic of Slovenia.

2. Methodology

Despite many frameworks related to performance and activity monitoring (Data-driven Decision Support System (DSS) (Power, 2002), Performance Monitoring, Business Performance Management (BPM), Business Activity Monitoring (BAM) (Dresner, 2003) etc.), there is a lack of methodologies for representing the concept of monitoring based on different data analysis methods. A similar methodology is addressing Performance Monitoring Protocols presented by the Working Party on Performance Monitoring in the Public Services (Bird, 2005). In this section, we present our approach to performance monitoring in modelling the PHC network in Slovenia.

2.1 Approach to human resources monitoring

Our model for monitoring the network of primary-care professionals in the Slovenian HCS consists of hierarchically connected modules. Each *module* is aimed at monitoring some aspect of the PHC network, which is of interest for decision-makers and managers of the network. Typical aspects about physicians are, for example: age and qualification of physicians, their workload and geographical distribution.

Each module involves a number of monitoring processes, which are gathered according to a given monitoring goal. Each *monitoring process* is characterised by: monitoring objectives, input data, data collecting methods, constraints on data, data dimensions, data analysis methods, output data, target criteria or target values of outputs, output data representation and visualisation methods, security requirements and the users of the monitoring system. Among these components, the *data analysis methods* transform the *input data* to *output data* represented using some *data representation formalism* according to the given *monitoring objectives*. The *target* is a level of performance that the organization aims to achieve for a particular activity. Information about *data collection* shows how and how often the data has been collected or needs to be collected (e.g., data can be collected by representative surveys or by standard procedures in organizations according to some refreshment rate). The *constraints* define the valid input and output data. *Security requirements* define the use and management of the monitoring processes and of the data.

This approach is not limited to any particular *data analysis method*. In principle, any methods can be used, such as Structured Query Language (SQL) procedures, On Line Analytical Process (OLAP) techniques for interactive knowledge discovering, as well as knowledge discovery in data (KDD) and data mining methods (Han, 2001) for discovering important but previously unknown knowledge. The same holds for *data representation* methods, which can include pivot tables, charts, network graphs and maps.

With respect to monitoring goals, *output variables* can be classified in different categories like *lead* and *lag* (Niven, 2003). The *lead* ones measure the performances that have influence on achieving the goals, whereas the *lag* are only related to the degree of achieving the goals.

In order to improve the comprehensibility of the model, its modules are hierarchically structured. The modules at the top level represent the main objectives. Usually all the main objectives can be incorporated in a single top-level module. The modules at a lower level are connected to the one at a higher level. Each connection represents a data channel that connects outputs of the lower level module with the inputs of a higher-level module. In principle, the hierarchy is constructed so that the results of lower-level processes could help to explain the results of monitoring processes at a higher level. For example, the module for the assessment of HCS responsiveness could be composed of the physical accessibility of Health Services, availability of resources of Health Services and the rate of visits of population to health-care provider.

2.2 The model of monitoring human resources in a HCS

The model of monitoring the network of physicians at the PHC level is made in accordance with the above described methodology. The real challenge was to improve the monitoring of human resources in the HCS by different KDD methods..

The main concept is described by the hierarchically connected modules, shown in Fig. 1. The module *human resources* represents the main aspect of monitoring. The included monitoring processes are aimed at the monitoring of anomalies, outliers and other interesting events related to the network of physicians. The lower-level modules intend to provide detailed explanations of these events.

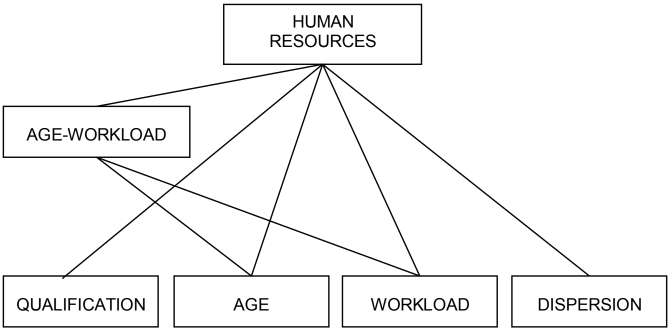


Fig. 1. The structure of the model for monitoring the network of physicians at a primary health-care level, represented by hierarchically connected modules.

3. Description of individual HCS modules

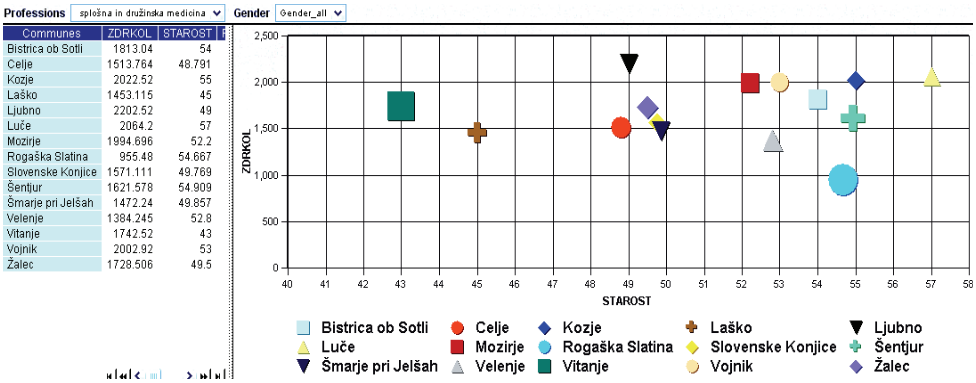
3.1 Monitoring of human resources

The main module *human resources* is aimed at a holistic monitoring of physicians' performance. In principle, the monitoring processes in this module must be able to process several input parameters. Therefore, different methods of KDD and multi-criteria decision models can be used for data analysis. The module includes a monitoring process that intends to represent the main aspects of physicians characterised by their *qualification*, *age*, *gender*, *workload* and *dispersion* (the number of locations where the physician works). The monitoring process is based on the OLAP model, which uses the dimensions: *time*, *location* where physicians work, *specialisation* and *gender*. The results based on a prototype application are presented by pivot tables and multidimensional charts, as shown in Fig. 2.

The scatterplot on the right side of Fig. 2 shows the average age, average workload, and average dispersion of physicians in communities for different specializations. The x-axis shows the average age of physicians, while the average workload is shown along the y-axis.

The communities are shown by shapes and colours of data points as explained in the legend. The size of these points is proportional to the average dispersion of physicians in the community. The specialization and gender of physicians could be selected using combo boxes in the top left corner. Thus, the scatterplot shows these average values for a gynaecologist working in some Slovenian community. For example, the workload and age of physicians shown on the top right corner are above the average. This presentation clearly shows outliers and anomalies in the HCS. Detailed information about these interesting aspects could be found in lower level modules. On the left side of Fig. 2, the same data is represented by the pivot table.

This module also includes a monitoring process aimed at discovering relations between the main aspects of physicians using the methods of association rules discovery (Srikant, 1996). The monitoring process tries to find the outliers overlooked by previous OLAP analyses. Table 1, for example, includes some rules focused on the relations between communities and physicians working in general practice. For example, for the community Škofja Loka it is characteristic that physicians younger than 40 years are underloaded (see the rules 2, 3). This module could also include the monitoring processes based on multi-criteria decision models (Bohanec, 2006).



Professions: splošna in družinska medicina
Gender: Gender_all

Communes	ZDRKOL	STAROST
Bistrica ob Sotli	1813.04	54
Celje	1513.764	48.791
Kozje	2022.52	55
Laško	1453.115	45
Ljubno	2202.52	49
Luče	2064.2	57
Mozirje	1994.696	52.2
Rogaška Slatina	955.48	54.667
Slovenske Konjice	1571.111	49.769
Šentjur	1621.578	54.909
Šmarje pri Jelšah	1472.24	49.857
Velenje	1384.245	52.8
Vitanje	1742.52	43
Vojnik	2002.92	53
Žalec	1728.506	49.5

Fig. 2. The holistic aspect of physicians presented by OLAP techniques.

Rule	Supp.	Conf.	Lift
[age:60+]+[workload: middle] ==>Ptuj	0.51%	14.63%	5.99
[dispersion:1]+[age:to40]+[workload: small] ==>Škofja Loka	0.34%	5.97%	4.42
[age:to40]+[workload: small] ==> Škofja Loka]	0.34%	5.19%	3.85
[workload: large]+[age:40-60]+[gender] ==>Domžale	0.42%	7.04%	3.63
[age:to40]+[workload: middle] ==> Novo mesto	0.42%	9.62%	3.46
[gender:z]+[age:to40]+[workload: small]==> Domžale	0.34%	6.35%	3.27
[gender:z]+[age:to40]+[workload: middle]==> Novo mesto]	0.34%	8.89%	3.19
[workload: large]+[gender:m] ==>Domžale	0.51%	6.00%	3.09
[age:60+]+[workload: large] ==> Maribor	0.67%	25.81%	3.00
[age:60+]+[workload: large]+[gender:m]==> Maribor	0.59%	25.00%	2.90

Table 1. Association rules showing relations between communities and general practice physicians.

3.2 Qualification of physicians

The aim of this module is to enable monitoring of physicians' and dentists' qualification for the job they actually perform. The main performance indicator is the physician's specialization degree, granted by the Slovenian Medical Chamber, which must be verified every 7 years. The specialization degree is a prerequisite for getting a license for employment in a certain area of medicine.

To monitor the suitability of physicians for the job they are performing we have used social network visualization technique available in the social network analysis program named Pajek ("Spider" (Batagelj, 2006)). The monitoring of physicians' suitability is achieved by the monitoring of three variables: SPEC (specialization), LIC (luicence), and OPR (the type of patients that the physician is in charge of, categorized by patient type). The motivation for this analysis is based on the observation that physicians with a certain specialization may get different luicenses, and while a certain licence assumes that the physician will only deal with patients of a certain patient category, in reality she may be in charge of different types of patients (e.g., a paediatrician may provide health services to grown up patients, although she has a specialization in paediatrics and a licence in paediatrics).

It has also been observed that an individual physician may have several specializations, and several licences, and hence patients of different types. The Pajek diagram (Fig. 3) shows well the typical (thick lines – a high number of physicians) and atypical (thin lines – a low number of physicians) cases, which enable abnormality detection and further analysis of individual discovered anomalies.

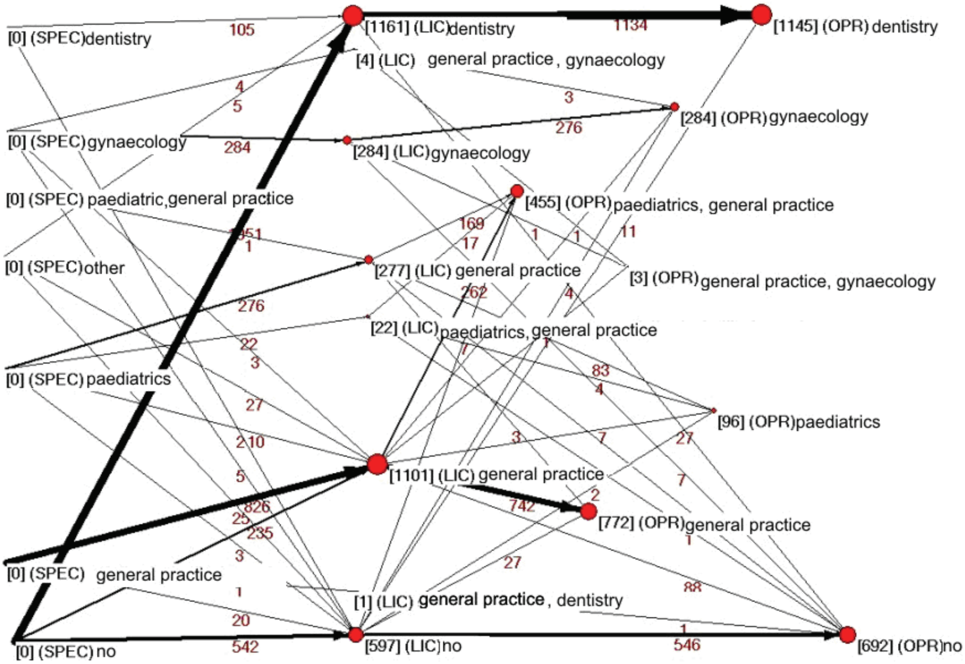

[0] (SPEC)dentistry
[0] (SPEC)gynaecology
[0] (SPEC) paediatric,general practice
[0] (SPEC)other
[0] (SPEC)paediatrics
[0] (SPEC) general practice
[0] (SPEC)no
[1161] (LIC)dentistry
[4] (LIC) general practice, gynaecology
[284] (LIC)gynaecology
[277] (LIC) general practice
[22] (LIC)paediatrics, general practice
[1101] (LIC) general practice
[1] (LIC) general practice, dentistry
[597] (LIC)no
[1145] (OPR) dentistry
[284] (OPR)gynaecology
[455] (OPR)paediatrics, general practice
[3] (OPR)general practice, gynaecology
[96] (OPR)paediatrics
[772] (OPR)general practice
[692] (OPR)no
105
1134
4
5
3
276
284
169
17
1
1
11
251
1
262
4
276
22
3
7
1
83
4
27
210
3
7
27
5
826
742
2
7
25
235
1
3
27
88
1
20
1
542
546

Fig. 3. The qualifications of physicians for the job they are performing.

3.3 Age of physicians

The monitoring processes in this module are aimed at age and gender analyses of physicians. The module includes a process based on the OLAP model. The dimensions of this model are *age*, *gender*, *specializations* and *locations* where they work. The main monitored quantity in the facts table is the number of physicians. The results are presented by pivot tables and different charts. For example, the number of gynaecologists by age and gender is presented in the chart in Fig. 4. The x-axis shows the age variable, while the number of doctors is shown along the y-axis. The gender is shown in the legend. Although relatively simple, this chart clearly shows a decline of the number of young male gynaecologists that have accomplished the studies in the last twenty years. Generally, the number of physicians is related to the number of students that have accomplished the studies at the Faculty of Medicine in Ljubljana that is the main source of doctors in Slovenia. Thus, this presentation can help planning the education system. The other process in this module provides a list of gynaecologists that are near retiring age. The list can be used to help planning missing human resources in the near future.

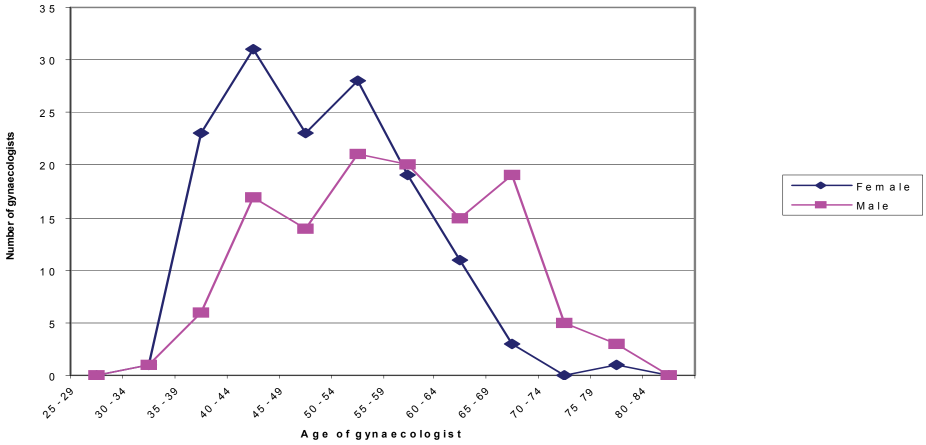


Fig. 4. The numbers of gynaecologists by age and gender.

3.4 Workload analysis

This module is aimed at monitoring the physicians' workload. Considering the available data, the assessment of workload is based on age-adjusted listed patients per physician. Each age group of listed patients is weighted according to use of health-care resources, e.g. the number of visits per physician. The physicians without registered patients are excluded from this analysis.

The monitoring process is based on the OLAP model. The dimensions of this model are: *time*, *specializations*, *locations* where the physicians work, the ratio between the number of listed patients and physicians, and the ratio between the age-adjusted number of listed patients and physicians. The measure in the fact table is the sum of physicians. Again, results are presented in pivot tables and different charts.

3.5 Age-workload analyses

The monitoring processes in this module are aimed at the combined analyses of physicians' age and their workload (number of patients per physician). The considered dimensions are: *time*, *locations* and *professions*. The monitoring process provides the state of each physician regarding their age and the number of patients. For example, Fig. 5 shows the age and workload of general practitioners. The x-axis shows the physicians' age, while the y-axis shows the number of their patients. Physicians aged between 40 and 50 have the most patients, but some of them have a large number of patients also after 50. The retirement of physicians having a large number of patients has an important impact on the PHC network. This impact is more precisely described by the next monitoring process, which is based on GIS.

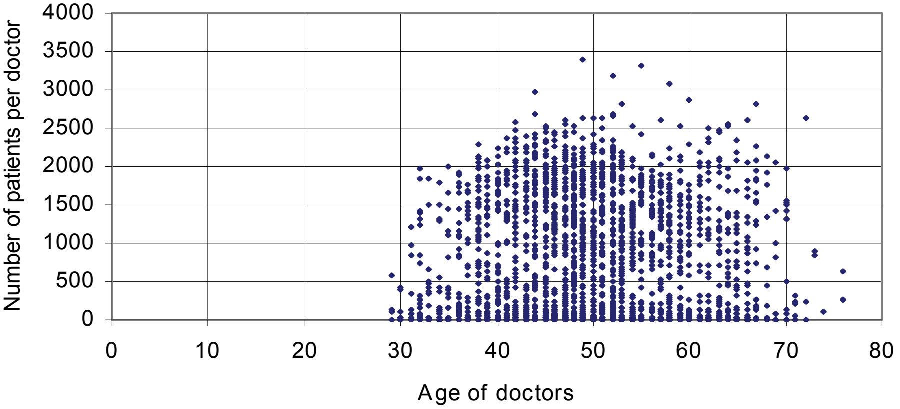


Fig. 5. The physicians in GP by age and listed patients.

The map in Fig. 6 shows the number and share of patients affected by the retirement of physicians in the next five years (until 2011). The assumption is that physicians will retire at the age of 65. In this map, each polygon represents a community. Their shade is proportional to the number of patients whose doctors will retire. The pie charts represent the ratio between the patients unaffected and affected by the retirement until 2011. Thus, this analysis provides information on the number of physicians and regions where they have to be replaced in the next five years.

From the implementation viewpoint, the latter module is composed of data about physicians' age, their registered patients and geographic data. The detailed information about physicians' age and patients is provided by subordinate modules *age* and *workload* (Fig. 1).

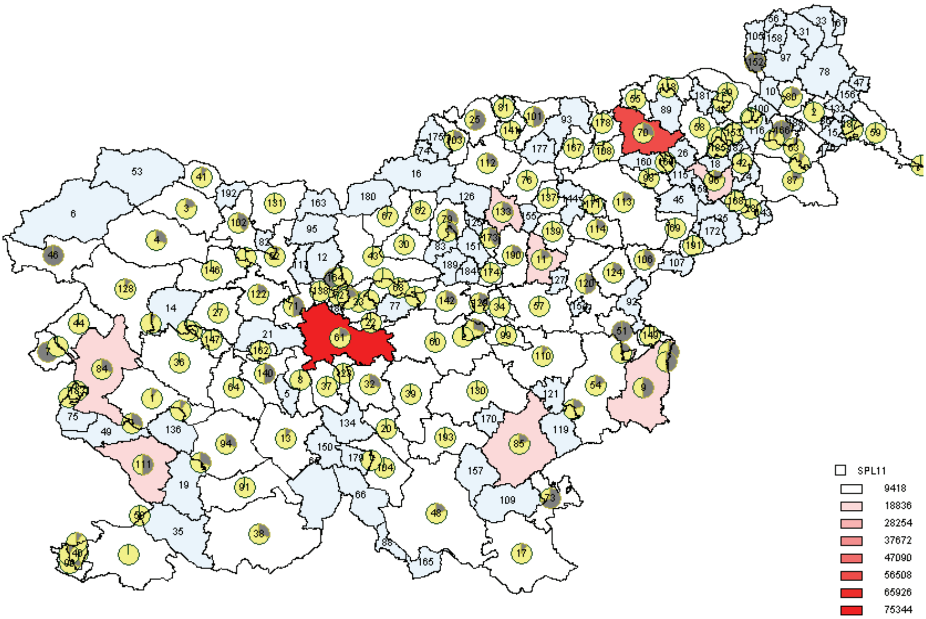


Fig. 6. The impact of physicians retiring until 2011.

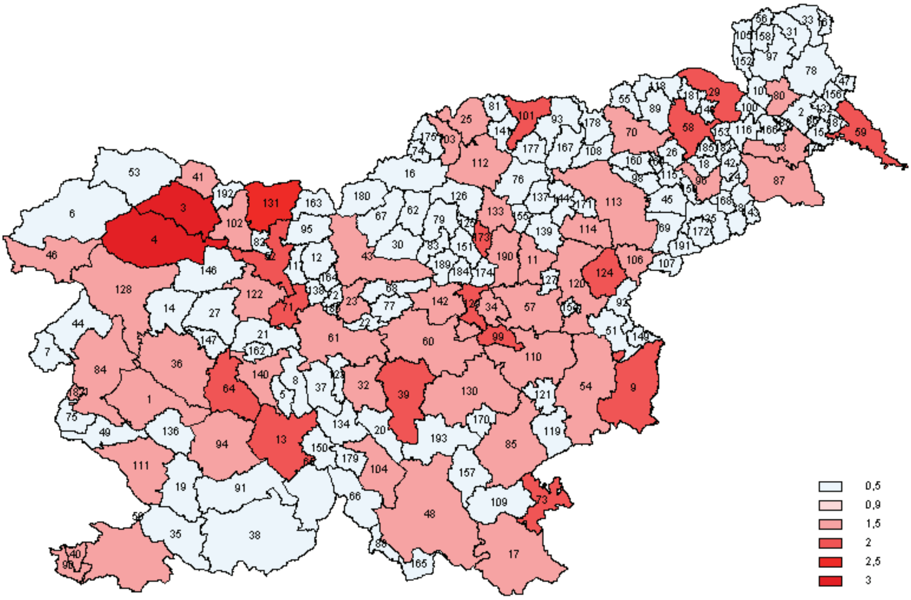


Fig. 7. Average dispersion of working locations.

3.6 Dispersion of working location

This module is aimed at monitoring the dispersion of locations where physicians work. Depending on the requirements, a physician may work on more than one location, but this dispersion usually means additional workload for physicians and their lower availability for patients at some location. The monitoring process provides the number of locations where physicians work, which are shown using the GIS techniques. Fig. 7, for example, shows the dispersion in 2005 for gynaecology. The darker the community are, the larger the average number of locations where physicians from this community work.

4. Conclusion

The aim of the presented human resource monitoring system is to assess performance and provide information necessary for the planning and management of primary health-care network in Slovenia. At the general level, the approach is based on a carefuly designed hierarchy of modules, each monitoring a specific aspect of the network – in particular physicians' age, qualifications, workload and dispersion. At the implementation level, the system uses a number of different techniques, including OLAP, KDD and data analysis (such as association rule mining). In most cases, the results are presented visually by charts, network graphs and maps. In this way, the monitoring system provides an information-rich picture of the network and its performance, and also helps detecting its critical aspects that require short- or long-term management actions. In principle, the higher levels of the model provide holistic information, while the lower levels provide more details that are useful for the explanation of observed phenomena.

The monitoring system has been developed in collaboration with the Ministry of Health of the Republic of Slovenia. Currently, it is implemented as a prototype and has been tested with real data for the year 2006. For the future, we wish that it becomes a regular tool for monitoring the health-care network in Slovenia. We also envision the application of the same methodology in other public networks, such as education and police.

5. References

World Health Organization: World Health Report 2000: Health Systems. Improving Performance, http://www.who.int/whr/2000/en/whr00_en.pdf, Accessed January 26, 2007 (2000)

Niven, R., P.: Balanced Scorecard for Government and Nonprofit Agencies, John Wiley and Sons, Inc (2003), ISBN 0-471-42328-9

Bird, M., S. (ed.): Performance indicators good, bad, and ugly, Working Party on Performance Monitoring in the Public Services, J. R. Statist. Soc. A (2005) 1-26

Bohanec, M.: DEXi, A Program for Multi-Attribute Decision Making. http://www-ai.ijs.si/MarkoBohanec/dexi.html, Accessed January 26, 2007 (2006)

Batagelj, V., Mrvar, A.: Program for Analysis and Visualization of Large Networks. Reference Manual, University of Ljubljana, Ljubljana (2006)

Dresner, H.: Business Activity Monitoring, BAN Architecture, Gartner Symposium ITXPO, Cannes, France (2003)

Power, J., D.: Decision Support Systems. Concepts and Resources for Managers, Quorum Books division Greenwood Publishing, ISBN: 156720497X (2002)

Han, J., Kamber, M.: Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers (2001)

Srikant, R., Agrawal, R.: Mining Quantitative Association Rules in Large Relation Tables, IBM Almaden Research Center, San Jose (1996)

6.3 Data Presentation Methods for Monitoring a Public Health Care System

This article describes some advanced data visualization techniques for the monitoring of the HCN (Pur, et al., 2007b). In accordance with the proposed methodology, these visualization techniques are aimed at the comprehensible visualization of results from various aspects in order to reveal their meaning. The techniques enable visual discovery of typical and atypical patterns, anomalies, and outliers in the health care data.

The article was presented at the 11th Mediterranean Conference on Medical and Biomedical Engineering and Computing 2007 MEDICON 2007, 26–30 June 2007, Ljubljana, Slovenia.

Data Presentation Methods for Monitoring a Public Health-Care System

Aleksander Pur[1], Marko Bohanec[2], Nada Lavrač[2,3], Bojan Cestnik[4,2]

[1] Ministry of the Interior, Ljubljana, Slovenia

[2] Jožef Stefan Institute, Ljubljana, Slovenia
[3] University of Nova Gorica, Nova Gorica, Slovenia
[4] Temida, d.o.o., Ljubljana, Slovenia

Abstract— **This paper present methods of data presentations that enable performance and activity monitoring of a health care system. The methods enable visual discovery of typical and atypical patterns, anomalies and outliers in the data. The methods were successfully implemented in a monitoring system developed for monitoring the primary health-care system of Slovenia, to be used by the national Ministry of Health.**

Keywords— **Data presentation methods, data visualization techniques, information graphics, Health Care System.**

I. Introduction

According to the World Health Report [1], a health-care system (HCS) is a system composed of organizations, institutions and resources that are devoted to producing a health action. A HCS contributes to good health, responsiveness to the expectations of the population, and fairness of financial contributions to health care [1]. The assessments of these contributions request the careful monitoring of performances and activities in HCS.

This paper focuses on the HCS of Slovenia which is divided into the primary, secondary and tertiary health-care levels. The primary health care (PHC) system is the patients' first entry point into the HCS. It is composed of four sub-systems: general practice, gynecology, pediatrics and dentistry. The paper illustrates the developed data presentation methods applied to performance and activity monitoring of the Slovenian HCS. These methods are included into the developed HCS monitoring model used at the primary health-care level, taking into account the physical accessibility to health care providers for patients, the availability of health care resources for patients and the rate of unregistered patients (the patients who have not have chosen their personal general practitioner) living in a certain area (community/region of Slovenia). This application was commissioned by the Ministry of Health of the Republic of Slovenia, who needs a holistic overview of the primary health-care network in order to make management decisions and apply appropriate management actions, as well as evaluate PHC target achievement. The term "data presentation" used in this paper in its broad sense includes - mostly visual - presentations of both single data elements and the presentation of more complex patterns, and it does not distinguish between "data", "information" "pattern" and "knowledge".

II. Method

Our approach to HCS monitoring is based on a model composed of hierarchically connected modules. Each module is aimed at monitoring a particular aspect of the HCS, which is of interest for decision-makers and managers of the system. Typical aspects about HCS are, for example, the accessibility to providers for patients, the qualification of physicians, their workload and their geographical distribution. Each module involves a number of monitoring processes, which are gathered according to a given monitoring goal. Each monitoring process includes one or more methods for data presentation; the same output data can be presented by different methods. Besides the output data presentation methods, each monitoring process is characterized by the monitoring objectives, input data, data collection methods, constraints on the data, data dimensions, data analysis methods, output data, target criteria or target values of outputs, security requirements and the users of the monitoring system. Among these components, the data analysis methods transform the input data to output data presented by some data presentation methods according to the given monitoring objectives.

This approach is not limited to any particular data presentation or analysis method. In principle, any data presentation method can be used, such as pivot tables, charts, network graphs and maps. The same holds for data analysis methods, which can include Structured Query Language (SQL) procedures, On Line Analytical Process (OLAP) techniques for interactive knowledge discovering, as well as knowledge discovery in data (KDD) and data mining methods [2] for discovering important but previously unknown knowledge.

The approach used in the HCS monitoring model is appropriate for hierarchically organized data presentations, as - in order to improve the comprehensibility of the model - the HCS modules are hierarchically structured. The modules at the top level represent the main monitoring processes/activities. The modules at a lower level are connected to a particular module at a higher level. Each connection represents a data channel that connects outputs of the lower-level modules with the inputs of a higher-level module. In principle, the hierarchy is constructed so that the output data of lower-level processes can help to explain the output data of monitoring processes at a higher-level. This functionality can be provided by the systems of menus, buttons and icons included in data presentation screens. The system helps the user to move from data presentation of higher-level monitoring process to appropriate data presentation of lower-level process or vice versa.

III. Short review of the HCS data

The central component of the HCS monitoring system is a data warehouse, which is composed of unique database entries from the following existing sources:

- Slovenian Social Security databases: the data about health-care providers together with assigned patients per individual general practitioner, the patients with social security, and the data about health care centers,
- the database of Slovenian physicians and dentists (provided by the Slovenian Medical Chamber),
- the database of the National Institute of Public Health containing data about Slovenian health centers, and
- the database of the Slovenian Statistics Bureau concerning the demographic and geographic distribution of citizens and communities in Slovenia.

This data warehouse contains real HCS data for the year 2006.

IV. Data presentations methods included in the HCS monitoring model

Each monitoring processes of our HCS monitoring model include one or more data presentation methods that present quantitative, ordinal or nominal data. In accordance with the object-oriented taxonomy of medical data presentations [3] these methods are divided into five major classes: list, table, graph, icon, and generated text, which can be further divided into eight subclasses as follows: a list presents text arranged in a one-dimensional sequence, a table presents items arranged in an n-dimensional grid, a graph is a spatial arrangement of point and lines with respect to axes conveying information, an icon presents a small stylized pictorial symbol and a generated text is related to computerized creation of text from coded data.

The data presentation methods in our HCS monitoring model are divided in accordance to data organization into three categories: text only, tables and information graphics. Information graphics - or “infographics” - are visual representations of information, data or knowledge, such as a graph, chart, flowchart, diagram, map and signage systems. The presentation methods are also divided in accordance with the used techniques into static and dynamic. In this paper the term static is aimed at data presentations that could be spread on the paper without data loss. On the contrary, the dynamic presentations provide full functionality only on a computer screen. Thus a typical static graph is presented as a picture without interactive functions and a typical dynamic graph is presented by OLAP techniques with drill-down, slice and dice functions. Some of these data presentations used in HCS monitoring model are described in this section. The model includes a lot of different data presentation methods, such as multidimensional maps, histograms and diagrams, but because of size limitation this paper is limited only to the four methods described below.

A. Simple graph-based static presentation of physicians’ qualifications

The aim of this data presentation is to enable monitoring of physicians’ and dentists’ qualification for the job they actually perform. The main performance indicator is the physician’s specialization degree, granted by the Slovenian Medical Chamber, which must be verified every 7 years. The specialization degree is a prerequisite for getting a license for employment in a certain area of medicine.

To monitor the suitability of physicians for the job they perform we have used a social network visualization technique available in the social network analysis program Pajek (“Spider” [4]). The monitoring of physicians’ suitability is achieved by the monitoring of three variables: SPEC (specialization), LIC (license), and OPR (the type of patients that the physician is in charge of, categorized by patient type). The motivation for this analysis is based on the observation that physicians with a certain specialization may get different licenses, and while a certain license assumes that the physician will only deal with patients of a certain patient category, in reality she may be in charge of different types of patients (e.g., a pediatrician may provide health services to grown up patients, although she has a specialization in pediatrics and a license in pediatrics).

The Pajek diagram (Fig. 1) shows well the typical (thick lines – a high number of physicians) and atypical (thin lines – a low number of physicians) cases, which enable abnormality detection and further analysis of individual discovered anomalies.

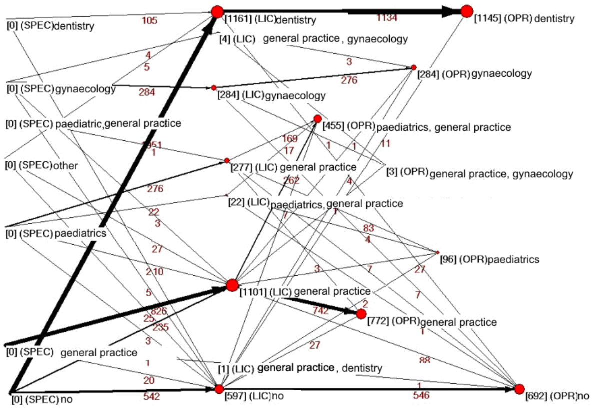
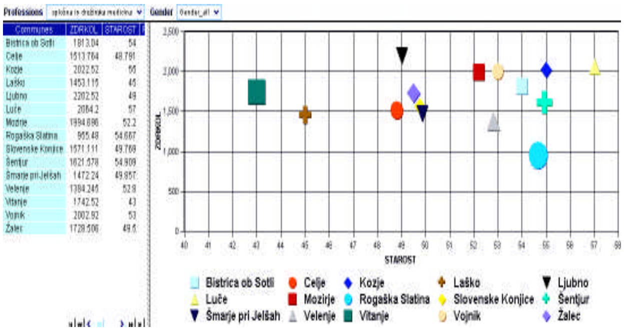

[0] (SPEC)dentistry
105
[1161] (LIC)dentistry
1134
[1145] (OPR) dentistry
[4] (LIC) general practice, gynaecology
4
5
3
[284] (OPR) gynaecology
276
[0] (SPEC)gynaecology
284
[284] (LIC)gynaecology
[0] (SPEC) paediatric,general practice
[455] (OPR) paediatrics, general practice
169
17
1
1
11
151
1
[0] (SPEC)other
[277] (LIC) general practice
[3] (OPR) general practice, gynaecology
262
4
276
[22] (LIC) paediatrics, general practice
22
3
7
1
83
4
[0] (SPEC) paediatrics
27
[96] (OPR)paediatrics
210
3
7
27
[1101] (LIC) general practice
5
7
826
25
235
742
2
[772] (OPR) general practice
1
3
27
[0] (SPEC) general practice
1
[1] (LIC) general practice, dentistry
88
20
1
[0] (SPEC) no
542
[597] (LIC) no
546
[692] (OPR) no

Fig 1. The qualifications of physicians for the job they are performing

B. Dynamic presentation based on OLAP techniques

Some data presentations implemented in the HCS monitoring model are based on OLAP techniques, as shown in Fig. 2. The scatter plot at the right-hand side of Fig. 2 shows the average age, average workload, and average dispersion of physicians in communities for different specializations. The x-axis shows the average age of physicians, while the average workload is shown along the y-axis. The communities are shown by shapes and colors of data points as explained in the legend. The size of these points is proportional to the average dispersion of physicians in the community. The specialization and gender of physicians could be selected using combo boxes in the top left corner. At the left-hand side of Fig. 2, the same data is represented by the pivot table.


Professions: splošna in družinska medicina
Gender: Gender_all

Communes	ZDRKOL	STAROST
Bistrica ob Sotli	1813.04	54
Celje	1513.764	48.791
Kozje	2022.52	56
Laško	1453.116	45
Ljubno	2202.52	49
Luče	2084.2	57
Mozirje	1994.896	52.2
Rogaška Slatina	955.48	54.667
Slovenske Konjice	1571.111	49.769
Šentjur	1621.578	54.909
Šmarje pri Jelšah	1472.24	49.857
Velenje	1384.246	52.8
Vitanje	1742.52	43
Vojnik	2002.92	53
Žalec	1728.506	49.6

Fig 2. The holistic aspect of physicians presented by OLAP techniques

This multidimensional visualization clearly shows outliers and anomalies in the HCS. Considering the hierarchical design of the HCS monitoring model, the detailed information about outliers and anomalies can be discovered by lower-level data presentations.

C. Tabular and textual presentation of association rules

The HCS monitoring model includes also monitoring processes based on association rules discovery techniques aimed at discovering interesting relations between items in the health care data [5]. These rules can be presented by tables as shown in Table 1 that includes selected rules focused on the relations between communities and physicians working in general practice. For example, the community Kranj is characterized by the physicians younger than 40 years that are under loaded (see the rules 4).

These rules can also be presented as a computer generated text, such as (rule 4): *5.97% of man dentists age up to 40 years that workload small is working in the Škofja Loka community that is 2 times more than expected.*

Our opinion is that these methods can be used with OLAP-based visualizations because some interesting relations omitted by OLAP techniques can be found without slice/dice, drill down and up activities.

Table 1. Association rules showing relations between communities and dentistry.

Rule	Supp.	Conf.	Lift
[workload: large] ==>Rogaška Slatina	0.21%	8%	19.3
[age:60+]+[workload: small]+ [gender:M] ==>Maribor	0.62%	20%	2.4
[age:60+]+[workload: small]+ [gender:M] ==>Murska Sobota	0.31%	5%	2.5
[age:to40]+[workload: small]+ [gender:M] ==>Kranj	0.41%	5.97%	2

D. Dynamic presentation of association rules

Usually, association rules are presented by texts and/or tables and arranged according to the parameters, such as support, confidence and lift. When we try, for example, to find associations between the number of visits to HCS providers and age-gender grouped patients from communities, or associations between illnesses caused job absences of age-gender based groups of individuals working in certain industrial branches and their illness we deal with a limited number of association rules with a predefined structure. A clear visualization of all these rules can be achieved by a matrix of rectangles association rules presentation where each association between two sets is presented by a rectangle (Fig 3). The size and color of rectangles depends of the association rules parameters support and lift. Thus, the size of a rectangle depends of the support parameter (the number of items included in the column and row), and the color of the rectangle depends of the lift parameter (the ratio of confidence to the expected confidence) [5].

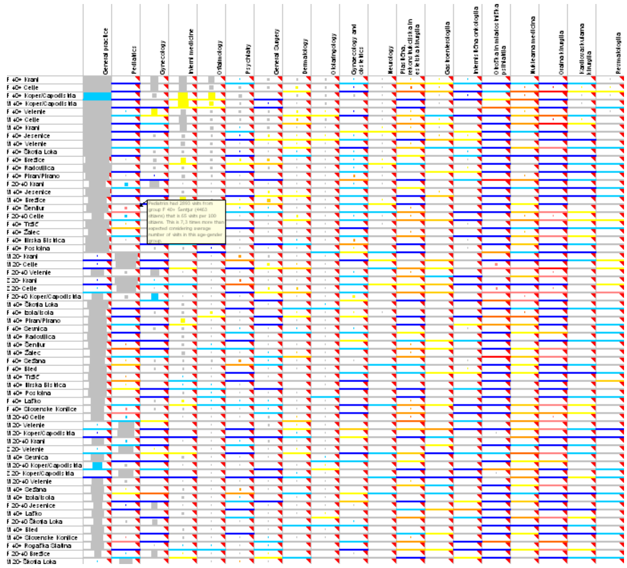
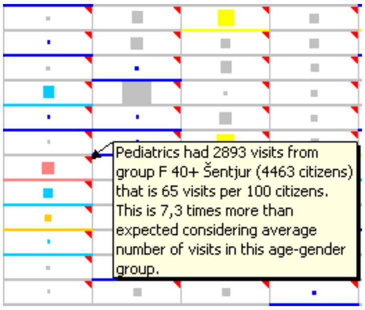


Fig 3. Association rules presented by a matrix of rectangles

A detailed information about each association can be shown by moving the mouse, thus moving the screen pointer on the selected rectangle (Fig 4). For example, the matrix in Fig. 4 shows the relations between the visits to a HCS provider of age-gender characterized patient group and the communities in which they live. The red rectangles present strong relations (with lift higher than 4), and the size of rectangles depends of the number of visits to certain HSC providers (the selected column) of gender-age grouped patients from these communities (the selected row).



Pediatrics had 2893 visits from group F 40+ Šentjur (4463 citizens) that is 65 visits per 100 citizens. This is 7,3 times more than expected considering average number of visits in this age-gender group.

Fig 4. Detailed information about a selected association

A detailed explanation of each relation can be listed in the message box (Fig 4). This kind of visualization enables us to visualize a large number of relations on a single screen. Therefore this presentation enables that some interesting relations can be found at a glance.

V. Conclusion

This paper describes four of the developed methods of data presentations that enable performance and activity monitoring in a HCS. Some of them can be used for discovering typical and atypical cases, such as the visualization of physicians' qualifications by the Pajek diagram (Fig. 1), discovering the anomalies by a matrix of rectangles (Fig. 3) and discovering of outliers by multidimensional graphs (Fig. 2). These methods are included in a HCS monitoring model made for the primary health-care level of Slovenia, while conceptually the model is not limited to any particular data presentation and analysis method. Our experiments have also proven the utility of the hierarchically designed HCS monitoring model, as it enables the user to track down interesting (unusual or unexpected) processes and activities in a HCS.

Acknowledgment

We gratefully acknowledge the financial support of the Slovenian Ministry of Education, Science and Sport, and the Slovenian Ministry of Health which is the end-user of the results of this project.

References

1. World Health Organization (2000) World Health Report 2000: Health Systems. Improving Performance, Http://www.who.int/whr/2000/en/whr00_en.pdf, Accessed January 26, 2007
2. Han J, Kamber M (2001) Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers
3. Starren J, Johnson S (2000) An Object-oriented Taxonomy of Medical Data Presentations. Journal of the American Medical Informatics Association Volume 7 Number 1 Jan / Feb 2000
4. Batagelj V, Mrvar A (2006) Program for Analysis and Visualization of Large Networks. Reference Manual, University of Ljubljana, Ljubljana
5. Srikant R, Agrawal R (1996) Mining Quantitative Association Rules in Large Relation Tables, IBM Almaden Research Center, San Jose

6.4 Data Mining and Visualization for Decision Support and Modeling of Public Health Care Resources

This article is focused on the detection of the similarities among community health centers in the Celje region, and on the measuring and visualization of the availability and accessibility of various public health care resources (Lavrač, 2007).

The goal of the monitoring of the availability and accessibility of the health care providers was to detect the municipalities that are health care underserved. Since the migrations of patients into other municipalities are an important factor, we proposed a novel measure for monitoring of the availability of the health care providers, which is described in the article. The results were presented through accessibility and availability maps and a scatter plot. The article was published in the *Journal of Biomedical Informatics*.

Available online at www.sciencedirect.com

ScienceDirect

Journal of Biomedical Informatics 40 (2007) 438–447

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Data mining and visualization for decision support and modeling of public health-care resources

Nada Lavrač [a,b,*], Marko Bohanec [a], Aleksander Pur [c], Bojan Cestnik [a,d], Marko Debeljak [a], Andrej Kobler [e]

[a] *Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia*
[b] *University of Nova Gorica, Nova Gorica, Slovenia*
[c] *Ministry of Interior Affairs, Štefanova 2, Ljubljana, Slovenia*
[d] *Temida, d.o.o. Ljubljana, Slovenia*
[e] *Slovenian Forestry Institute, Ljubljana, Slovenia*

Received 1 July 2006
Available online 10 November 2006

Abstract

This paper proposes an innovative use of data mining and visualization techniques for decision support in planning and regional-level management of Slovenian public health-care. Data mining and statistical techniques were used to analyze databases collected by a regional Public Heath Institute. We also studied organizational aspects of public health resources in the selected Celje region with the objective to identify the areas that are atypical in terms of availability and accessibility of public health services for the population. The most important step was the detection of outliers and the analysis of availability and accessibility deviations. The results are applicable to health-care planning and support in decision making by local and regional health-care authorities. In addition to the practical results, which are directly useful for decision making in planning of the regional health-care system, the main methodological contribution of the paper are the developed visualization methods that can be used to facilitate knowledge management and decision making processes.

Keywords: Data mining; Decision support; Knowledge discovery; Knowledge management; Visualization; Applications to health-care

1. Introduction

Effective medical prevention and good access to health-care resources are important factors that affect citizens' welfare and quality of life. As such, these are important factors in strategic planning at the national level, as well as in planning at the regional and local community levels. Large quantities of data collected by medical institutions and governmental public health institutions can serve as a valuable source of evidence that needs to be taken into account when making decisions about priorities to be included in regional strategic health-care plans.

Slovenian regional public health institutes (PHIs), coordinated by the national Institute of Public Health (IPH), are an important part of the system of public health in Slovenia. Their functions are public health monitoring, organizing public health-related activities and proposing and implementing actions for maintaining and improving public health. PHIs themselves coordinate a regional network of hospitals, clinics, individual health professionals and other health-care resources involved in particular health-care activities. Data at all levels are collected, and a national-level data warehouse is maintained at the national IPH.

* Corresponding author. Fax: +386 1 477 3315.
E-mail addresses: nada.lavrac@ijs.si (N. Lavrač), marko.bohanec@ijs.si (M. Bohanec).

1532-0464/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2006.10.003

This paper describes an application of data mining and decision support in public health-care, carried out in Slovenia within a project called MediMap. The goal of MediMap was to improve health-care knowledge management through data mining and decision support integration [3]. *Data mining* [4] is concerned with finding interesting patterns in data. Data mining includes predictive data mining algorithms, which result in models that can be used for prediction and classification, and descriptive data mining algorithms for finding interesting patterns in the data, like associations, clusters and subgroups. Data mining is typically applied to knowledge discovery in large and complex databases and has been extensively used in knowledge management [1] and industrial and business problem solving [2]. On the other hand, decision support [5,6] is concerned with helping decision makers solve problems and make decisions. As indicated by the results of recent research [3], data mining and decision support integration can lead to improved solutions in practical applications.

Health-care is a knowledge-intensive domain, in which neither data gathering nor data analysis can be successful without using knowledge about both the problem domain and the data analysis process. This indicates the usefulness of integrating data mining with decision support techniques [3,5] to promote the construction of effective decision criteria and decision models supporting decision making and planning in public health-care. The integration of the data mining and decision support approaches, as well as the novel visualization techniques developed for the purpose of this health-care application, have facilitated knowledge management and improved decision support.

In MediMap, we mainly used descriptive data mining methods and combined them with visualization and multi-criteria decision support techniques to improve the management of data and knowledge at the Public Health Institute of the Celje region. The main objective of MediMap was to set up appropriate models and tools to support decisions concerning regional health-care, aimed to serve also as a reference model for other regional PHIs. We approached this goal in two phases: first, we analyzed the available data with data mining techniques, and second, we used the results of data mining for a more elaborate study using decision support techniques. In the first phase we focused on the problem of directing the patients from primary health-care centers to specialists. In the second phase we studied organizational aspects of public health resources in the Celje region with the goal to identify the areas that are atypical in terms of availability and accessibility of public health services.

The paper is organized as follows. Section 2 presents data mining and decision support used as the main technologies used for knowledge management in this application. Section 3 presents the data that was used for the analysis of the Celje health-care resources. The results of applying data mining and decision support techniques, and the visualization of the results, are presented in Section 4. Section 5 concludes by summarizing the main results and by presenting plans for further work.

2. Data mining and decision support for knowledge management

Data mining [3,4] is concerned with finding models and patterns from the available data. Data mining includes predictive data mining algorithms, which result in models that can be used for prediction and classification, and descriptive data mining algorithms for finding interesting patterns in the data, like associations, clusters and subgroups.

Decision support [3,5] is concerned with helping decision makers solve problems and make decisions. Decision support provides a variety of data analysis, preference modeling, simulation, visualization and interactive techniques, and tools such as decision support systems, multiple-criteria modeling, group decision support and mediation systems, expert systems, databases and data warehouses. Decision support systems incorporate both data and models.

Data mining and decision support can be integrated to better solve data analysis and decision support problems. In *knowledge management* [1], such integration is interesting for several reasons. For example, in data mining it is often unclear which algorithm is best suited for the problem. Here, we require some decision support for data mining. Another example is when there is a lack of data for the analysis. To ensure that appropriate data is recorded when the collection process begins it is useful to first build a decision model and use it as a basis for defining the attributes that will describe the data. These two examples show that data mining and decision support can complement each other, to achieve better results. Different aspects of data mining and decision support integration have been investigated in [3].

3. Public health data

To model the Celje regional health-care system, we first wanted to better understand the health-care resources and their connections in the Celje region. The location of this region on the map of Slovenia is shown in Fig. 1. The Celje region is composed of 11 communities, further divided into 34 local communities.

For the purpose of MediMap, data mining techniques were applied to the data of 11 community health centers (CHCs) of the Celje region. The dataset consisted of three databases:

- The health-care providers database,
- The out-patient health-care statistics database (patients' visits to general practitioners and specialists, diseases, human resources and availability), and
- The medical status database.

To model the processes of a particular CHC (the patient flow), we used additional data describing the directing of patients to other CHCs or specialists.

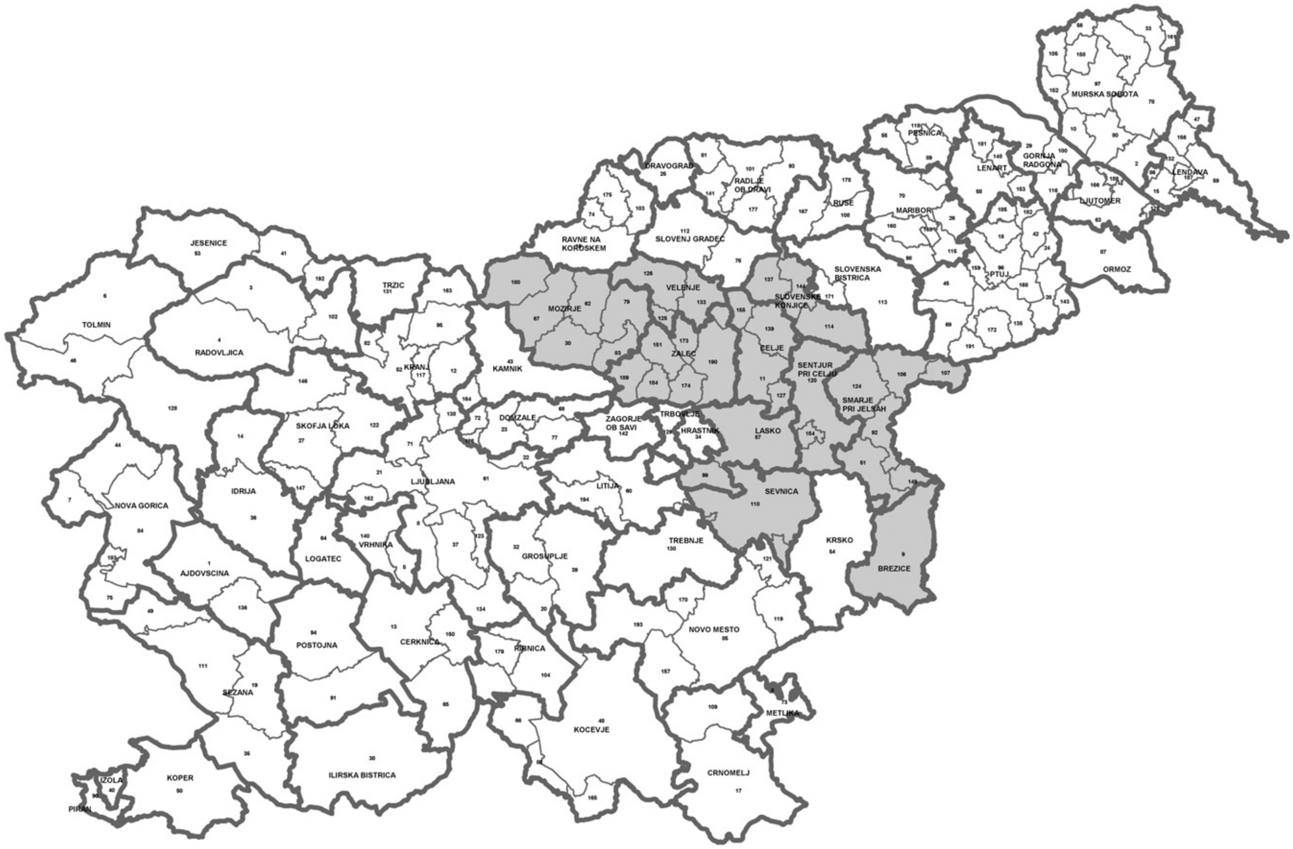


Fig. 1. The Celje region, located on the map of Slovenia.

4. Results of analyses

This section presents the detected similarities of community health centers of the Celje region, the analysis of the availability and accessibility of various public health-care resources, as well as the achieved results of decision support and visualization allowing for more advanced planning of health-care resources.

4.1. Detecting similarities of community health centers with data mining

The goals of this analysis were to detect the similarities between CHCs, and to detect the atypical CHCs. Similarities between CHCs were analyzed according to four different categories: (a) patients' age categories, (b) patients' social categories, (c) organization of the CHC and (d) employment structure of the CHC (Table 1). The categories (a)–(c) are described by five attributes and (d) is described by four attributes. The attributes of categories (a) and (b) are numeric and represent relative frequencies (e.g., value x of the attribute pre-school means that in a given CHC x% of patients are pre-school children).

For each category, similarity groups were constructed using four different clustering methods: agglomerative classification [7], principal component analysis [7], the Kolmogorov–Smirnov test [8], as well as the quantile range test and polar ordination [9]. An illustration of clusters, generated by Ward's agglomerative hierarchical clustering

Table 1
Description of categories and attributes used in analyzing the similarities between CHCs

	Categories	Attributes				
a	Patients' age	0–6 (pre-school)	7–19 (school)	20–49	50–64	⩾65
b	Patients' social status	Blue-collar workers	Farmers	Pensioners	Unclassified	Other
c	Organization of a CHC	Years of operation	Contacts per hour	Contacts per practice	Number of surgeries	Contacts per employee
d	Employment structure of a CHC	Education level	Time since professional exam	Time since first employment	Average age of employee	

```
Celje         -+-------+
Žalec         -+       +-----------+
Laško         ---------+           +---------------------------+
Velenje       ---------------------+                           I
Sevnica       -+-+                                             I
Šentjur       -+ +-----+                                       I
Radeče        ---+     +-------------+                         I
Mozirje       ---------+             +-------------------------+
Brežice       ---+-------+           I
Slov.Konjice  ---+       +-----------+
Šmarje        -----------+
```

Fig. 2. Results of hierarchical clustering of CHCs.

using the Euclidean distance measure, is given in Fig. 2. According to the maximal inter-cluster dissimilarity, the methods splits the CHCs into two top-level clusters, Cluster 1 formed of upper four CHCs and Cluster 2 of the bottom seven CHCs.

The similarities of community health centers were presented and evaluated by PHI Celje domain experts. In several cases the results confirmed already known similarities, while the experts could not find obvious explanations of the results of clustering. To explain the main differences between clusters, we have transformed the result of clustering into a classification task, and used a decision tree learning algorithm (J48 WEKA implementation of the well-known C4.5 learner [10]) to get a decision tree distinguishing the two classes (the two top-level clusters). To illustrate the approach, take the two top-level clusters of Fig. 2, considered as two disjoint classes. Fig. 3 shows a decision tree in which only the most informative attribute, distinguishing between the two groups of health centers, is an attribute of category (a): the age of patients. Community health centers in which pre-school children (PreSc) constitute more than 1.41% of all visits to the center form Cluster 1 (consisting of seven health centers). The experts' explanation is that these centers lack specialized pediatrician services, hence pre-school children are frequently treated by general practitioners. This is undesirable and indicates the need for corrective health-care management decisions. Despite the simplicity of the presented result achieved, and the approach taken, the combination of clustering and decision tree learning turned out to be useful for achieving a better explanation of the results achieved, which were satisfactory to the health-care experts.

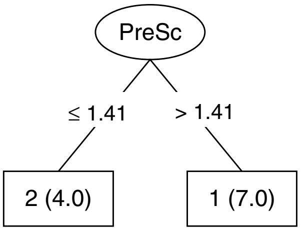


Fig. 3. A decision tree representation of the two clusters from Fig. 2, offering an explanation for the grouping into two classes (class 1 consisting of seven CHCs and class 2 consisting of four CHCs).

Averages over four clustering methods per category were used to further try to detect the similarities between the CHCs of the Celje region (Fig. 4). The results of these experiments confirmed some similarities between community health centers, but the similarity matrix did not provide novel explanations to the experts.

To further analyze the differences between the health centers, we have developed a different visualization method, enabling the analysis of the typicality of CHCs based on the comparison of the estimated number of patients that can be handled by a CHC (its capacity estimated by the number of employed staff) and the actual number of patients handled by the CHC. The outcome, shown in Fig. 5, was very much appreciated by the experts. The figure presents some atypical CHCs, deviating from the diagonal line, such as CHC Brežice and Žalec, which have insufficient staff compared to the number of actual patients requiring health-care services.

In summary, the results of these experiments confirmed some similarities between community health centers and pointed out atypical community health centers together with their properties that required corrective management activities. This part of the analysis, enabling decision support through the visualization of deviating/atypical CHCs, turned out to be most appreciated by the collaborating experts.

4.2. Availability and accessibility of public health-care resources

The goal of this analysis was to detect the local communities that are underserved concerning general practice health services—this means that the population in these areas has less than a generally accepted level of services available for the population. We evaluated 34 local communities in the Celje region. The evaluation was based on the ratio of the capacity of health-care services available to patients from the community and the demand for these services by the population of the same area.

For this analysis, the following novel measures and criteria were proposed. In our case, the *capacity* of health-care services is defined as available time of health-care services for patients in the given community, and *demand* means the number of accesses to health-care services from patients

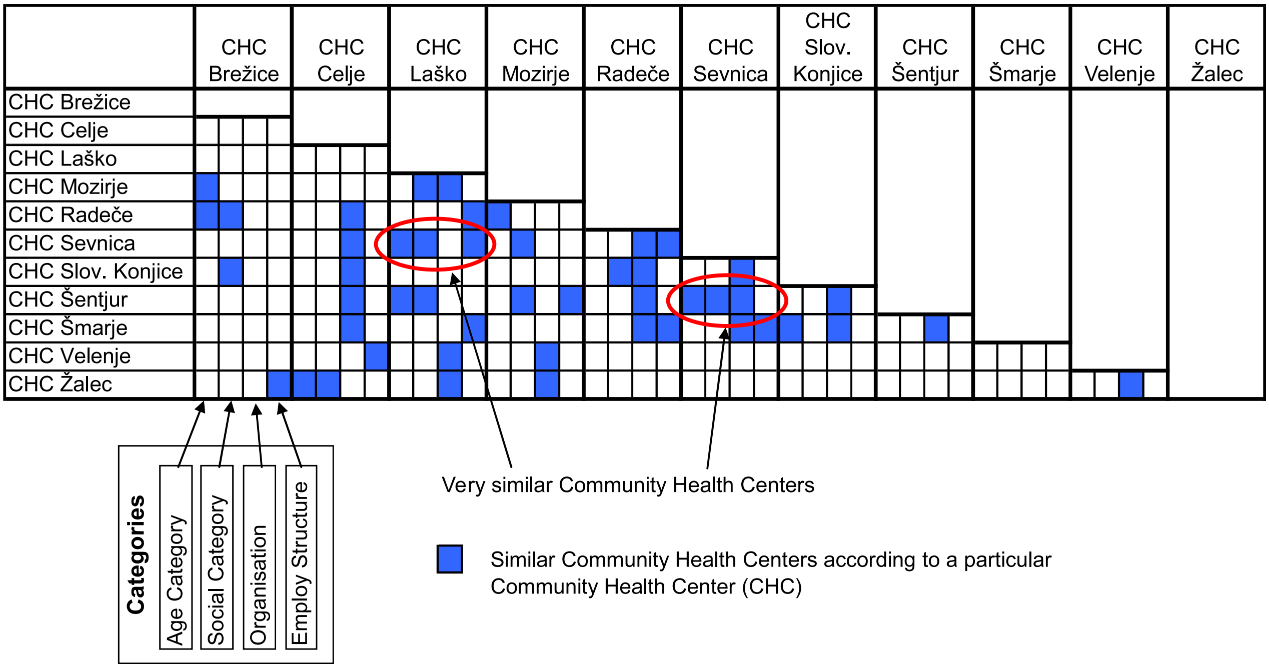


Fig. 4. The similarity matrix of community health centres in the Celje region.

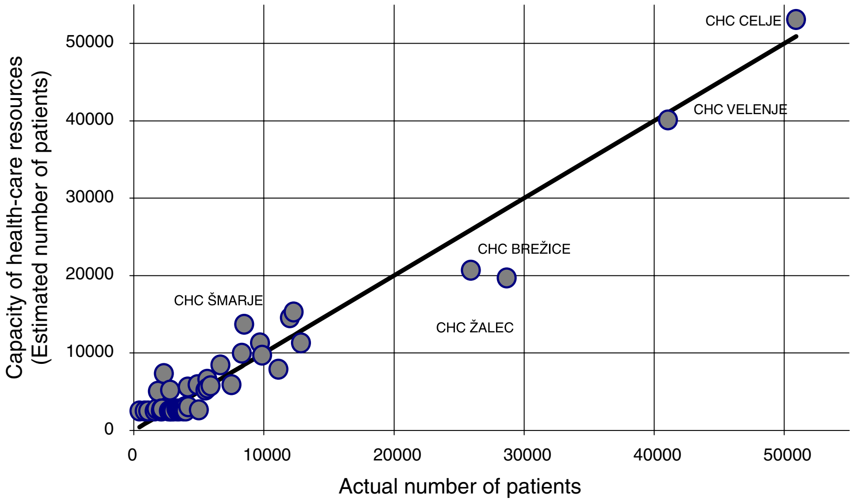


Fig. 5. Detecting atypical Celje region health-care resources (deviating from the diagonal line).

from the community. Therefore, our main criterion for the evaluation of the health-care system for patients in a community is actually the demand/capacity ratio, computed by the average time of available health services per access of a patient from the given community.

In the definition of this measure, called AHSP (availability of health services for patients)

$$\mathrm{AHSP} = \frac{\sum t_i}{p_c} \tag{1}$$

variable t_i denotes the total working time of health-care service i in community c, and p_c the number of accesses to health-care services of patients from community c. By setting the appropriate expert-defined threshold, this measure can turn into a criterion that can be used for decision support.

Notice that the AHSP measure does not take into account that many patients access health services in neighbouring or even more distant communities. Moreover, some of communities do not have their own health-care services at all. Since the migrations of patients into other communities are an important factor, we proposed a novel measure, $\mathrm{AHSP_m}$, which takes migrations into the account. $\mathrm{AHSP_m}$ is an average of available time per access of a patient in all of the health services, depending on the amount of patients from community c that each service received. First, we defined two variables: X_c—the available time per access of a patient from community c and Y—the health service (which can have values 1, 2, 3,...). We got the desired average with the help of the law of total expectation:

$$E(X_c) = \sum_i E(X_c|Y = i)P(Y = i) \tag{2}$$

The term $E(X_c|Y = i)$, which we denote by a_i, is the available time per access of a patient from community c at health-care service i. It can be calculated as the ratio of the total working time of health-care service and the total number of visits. The probability $P(Y = i)$ that the patient visited health-care service i can be stated as the ratio of the number of accesses of patients from community c to health service i (denoted p_{ci}), and the total number of accesses of patients from community c (already defined as p_c). Consequently, we can write the new criterion as

$$\mathrm{AHSP_m} = \sum_i a_i \frac{p_{ci}}{p_c} = \frac{1}{p_c} \sum_i a_i p_{ci} \tag{3}$$

The evaluation of communities in the Celje region using the AHSP and $\mathrm{AHSP_m}$ measures is shown in Figs. 6 and 7, respectively. The color intensity represents the availability of health services for patients: the darker the color, the higher the health-care availability in the community (measured in hours per visit).

Notice the advantage of the modified measure proposed in Eq. (3): the main difference between the evaluations employing the two measures is namely noticeable in communities which do not have own health-care services, like Braslovče, Tabor, Dobje and Solčava. If the migrations of patients to neighbouring communities are not considered, then it looks as if the inhabitants of these communities were without access to health-care resources (Fig. 6). Therefore, $\mathrm{AHSP_m}$ (Fig. 7) is a more realistic evaluation measure and by appropriately setting the threshold values, can be turned into an appropriate criterion to be used by health-care decision makers. Having determined different threshold values, a geographical representation of the results was made possible. This result visualization was very well-accepted by the health-care experts.

To further refine the analysis concerning the availability of health-care services and for the purpose of its visualization (Fig. 8), we introduced two additional measures. The AHS (availability of health services) measure was defined, aimed at measuring the availability of health-care services for the population from a community. More precisely, AHS is defined as the available time of health-care services per population g_c from community c, considering the migrations:

$$\mathrm{AHS} = \frac{1}{g_c} \sum_i a_i p_{ci} \tag{4}$$

The next measure RAHS (rate of accesses to health services), defines the rate of accesses to health-care services for population g_c from community c:

$$\mathrm{RAHS} = \frac{p_c}{g_c} \tag{5}$$

In this case, $\mathrm{AHSP_m}$ is defined as the ratio between the availability of health services for population from the community and the rate of visiting the health services:

$$\mathrm{AHSP_m} = \frac{\mathrm{AHS}}{\mathrm{RAHS}} \tag{6}$$

All these measures, and the derived criteria based on threshold values, give us some very interesting indicators of health conditions and the availability of health-care services in different communities. Using a novel visualization method developed for this purpose, they can be

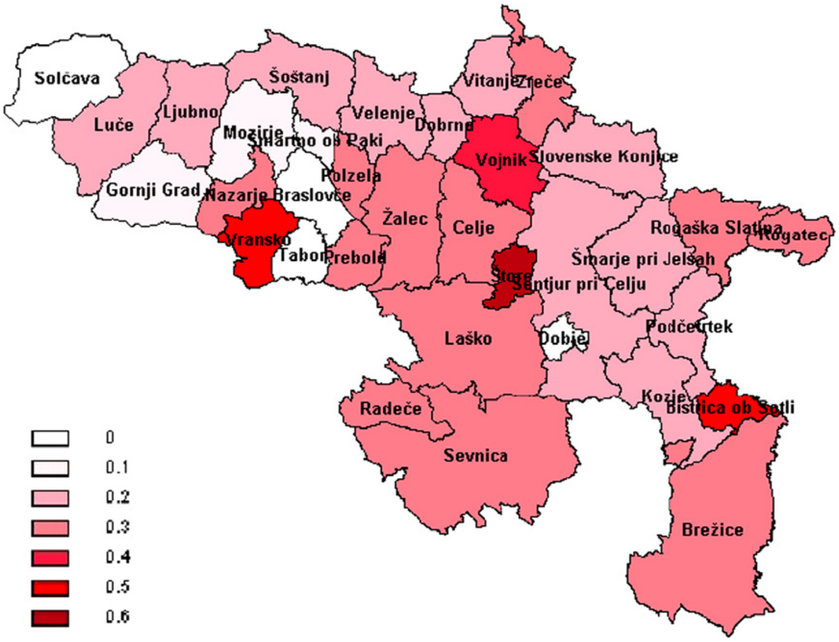


Fig. 6. Availability of health services (AHSP), measured in hours per visit, in the Celje region in 2003.

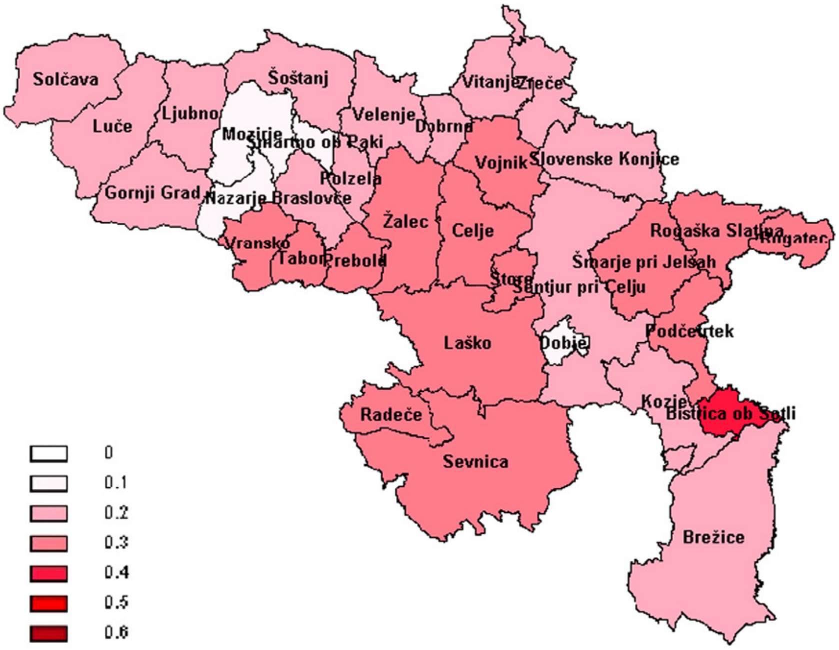


Fig. 7. Availability of health services in Celje in 2003, measured in hours per visit, considering the migrations of patients to neighboring communities ($AHSP_m$).

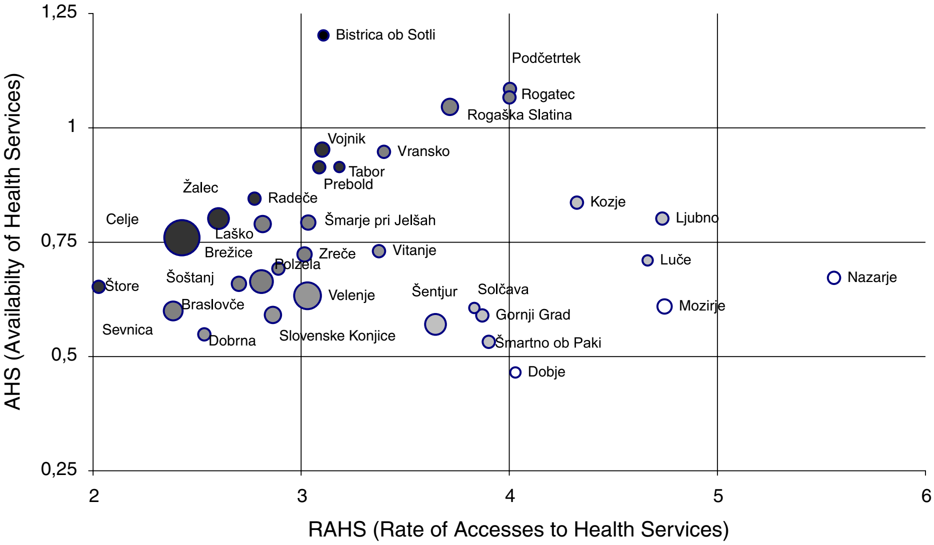


Fig. 8. Available time of health-care services per population by community.

conveniently presented as shown in Fig. 8. Four measurements are actually shown in the chart: RAHS along the horizontal axis, AHS along the vertical axis, $AHSP_m$ as dot color intensity, and population size g_c as dot diameter. Communities with average values of RAHS and AHS appear in the middle of the chart. The outliers represent unusual communities regarding health-care. Communities at the left side of the chart have lower rate of accesses to health services and the ones at the right side have higher access rates. Communities with lower values of AHS are located at the bottom, and those with higher values at the top. The dark-colored communities have higher values of $AHSP_m$ than the light-colored ones.

Consequently, by proposing a novel visualization of this multi-criteria problem, Fig. 8 enables the discovery of implicit and interesting knowledge about health-care services in different communities. For example, the reason for a high value of $AHSP_m$ in communities at the left side of the chart (e.g., Štore) could be the low rate of accesses to the nearest health services, caused by inappropriate medical procedures in these services. A possible reason for the low value of $AHSP_m$ in communities at the right side (Naz-

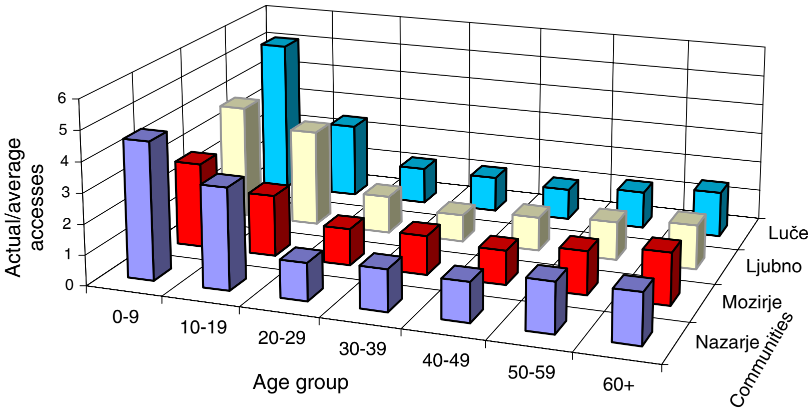


Fig. 9. The ratio between the actual and the average accesses to health services (in year 2003).

arje, Mozirje, Luče in Ljubno) might be high rates of accesses to health services. Further expert analysis was motivated based on this multi-criteria result visualization.

4.3. Decision support for planning health-care resources

Additional analysis of these rates can be provided by a chart shown in Fig. 9. The chart shows the ratio of actual rates of accesses of health services and expected rates for age groups of the population in the communities. This ratio is used in order to simplify the detection of unusual rates of accesses to health services. The expected rate of accesses to health services is the average rate of population in an age group. For example, the access to health services of the population aged between 0 and 9 years is almost five times more frequent than of the population aged between 20 and 29 years. The age group of population from communities is measured along the horizontal axis. Thus, the chart shows that in these communities the rate of accesses to health services of the population under 20 is unusually high. This finding motivated further analysis, which showed that the main reason for the high value of $AHSP_m$ in these communities is the absence of paediatric services, which was later confirmed by the health-care experts.

A further view on the disparity of health-care in the communities is provided in Fig. 10. There, the evaluation of health services is based on the ratio between the health-care capacity and demand. In our case the demand means the number of accesses to health services, and is measured along the horizontal axis. The capacity is proportional to the working time of health services, and is measured along the vertical axis. Some of the health services are denoted by an identification number and the community name. The regression line represents the expected working times of health services, with respect to the number of accesses. The working times of the health services under the regression line, like Nazarje and Mozirje, are too short, and of those above the regression line are too long. Consequently, this chart can serve for supporting decisions in planning the capacity and working times of health services. Methodologically, the aim of this chart is to highlight the CHCs that lie far away from the regression line rather than to accurately construct the regression line. In our case, we constructed the regression line using all the CHCs, without

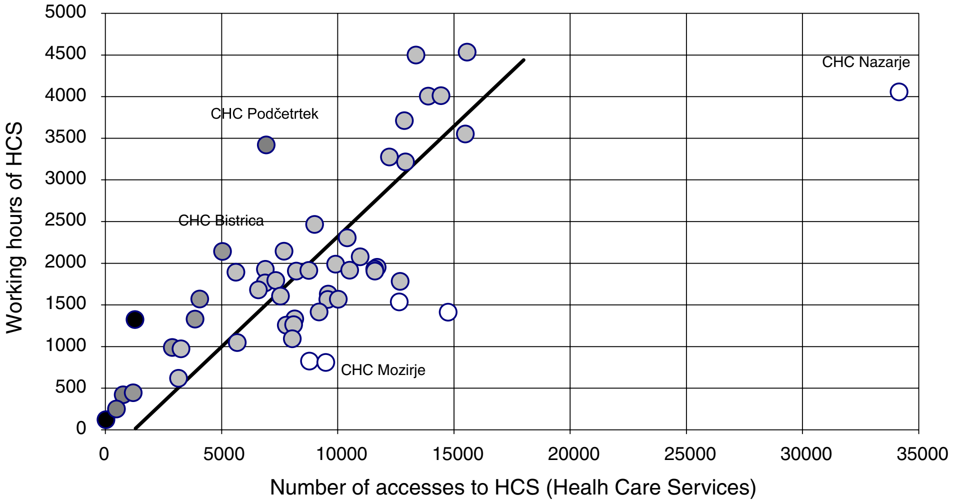


Fig. 10. Evaluation of health services: the ratio between the health-care capacity and the demand.

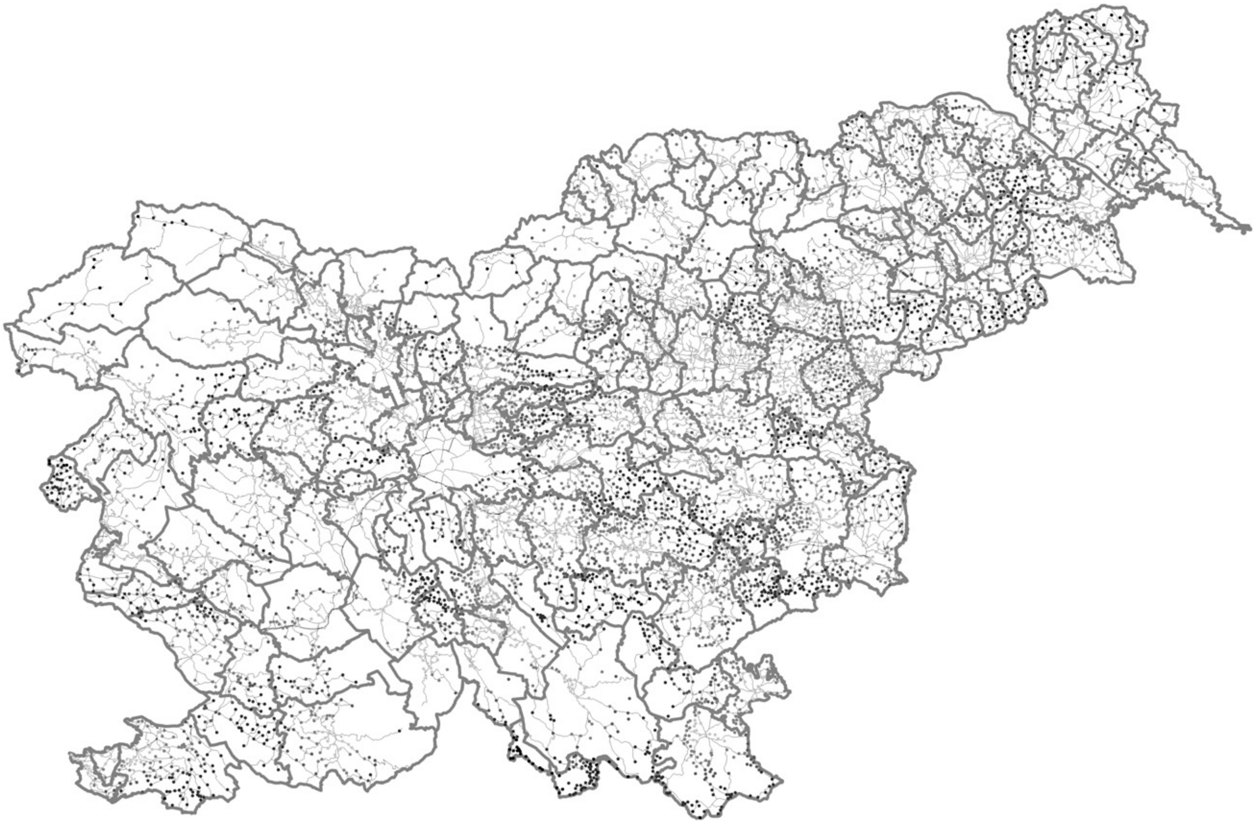


Fig. 11. CHC accessibility map for gynaecology for all the regions of Slovenia.

discarding any outliers as would be the case in the ordinary linear regression.

4.4. Decision support through GIS visualizations

GIS data can be used to visualize the national road network, detailed by road category, and the locations of community health centers of the map of Slovenia. Instead of the raw data visualization, presented by roads leading to the closest CHC for a citizen at a given location, we upgraded the road visualizations by computing the CHC accessibility through the so-called road "resistance" measure, which is anti-proportional to the average travel speed (which is—in turn—proportional to the road category). The following road categories, allowing for different access speeds, were taken into the account: highways (120 km/h), main roads (80 km/h), regional roads (60 km/h) and local roads (50 km/h). This lead to the development of the CHC "access" map, which enables the visualization of areas of Slovenia with low CHC access capacity. Such visualization enables the decision maker to see areas which have low accessibility to primary health services, possibly developing new health-care facilities in such regions. A sample access map for gynaecology for Slovenia is shown in Fig. 11. Each dot represents a settlement (town/village) and its intensity corresponds to the accessibility of the nearest gynaecological health service: the darker the dot, the lower the access capacity.

5. Conclusions

The use of data mining and decision support methods, including novel visualization methods, can lead to better performance in decision making, can improve the effectiveness of developed solutions and enables tackling of new types of problems that have not been addressed before. A real-life application of this approach in public health-care was shown in this paper, following some of the guidelines for public health management recommended in [11,12].

In the MediMap project we have developed methods and tools that can help regional public health institutes (PHIs) and the national Institute of Public Health (IPH) to perform their tasks more effectively. Tools and methods were developed for the reference case of PHI Celje and tested on selected problems related to health-care organization, accessibility of health-care services to the citizens and the health-care providers work. The main achievement was the creation of the model of the availability and accessibility of health services to the population of a given area. With the proposed model it was possible to identify the regions that differ from the average and to consequently explain the causes for such situations, providing many benefits for health-care planning and management processes.

In addition, the national IPH has used the results of this study to identify missing data that should be included in the improved protocol of public health data gathering at the national level, as the study indicated that additional—more detailed, but relatively easy to obtain—data from the community health centres was needed. This finding was valuable for the IPH, as this institution is in charge of defining the national data model and prescribing national data gathering rules and procedures.

In further work, we will extend this analysis to other regions of Slovenia. We will focus on the development of decision support tools for modeling of health-care providers using data mining. We wish to implement the developed methodology so that it can be regularly used for decision support in organizations responsible for the health-care network: the national Ministry of Health, the national IPH, and the regional PHIs.

Acknowledgments

We acknowledge the financial support of the Public Health Institute Celje and the Slovenian Ministry of Higher Education, Science and Technology. We also express our thanks to other members of the MediMap project team Tanja Urbančič, Mitja Jermol, Tadeja Kopač, Branko Kavšek and Primož Lukšič who also contributed to the results described in this paper.

References

[1] Smith RG, Farquhar A. The road ahead for knowledge management: an AI perspective. AI Magazine 2000;21(4):17–40.

[2] Biere M. Business intelligence for the enterprise. Engelwood Cliffs, NJ: Prentice Hall PTR; 2003.

[3] Mladenić D, Lavrač N, Bohanec M, Moyle S, editors. Data mining and decision support: integration and collaboration. Dordrecht: Kluwer; 2003.

[4] Han J, Kamber M. Data mining: concepts and techniques. 2nd ed. Los Altos, CA: Morgan Kaufman; 2006.

[5] Mallach EG. Decision support and data warehouse systems. New York: McGraw-Hill; 2000.

[6] Turban E, Aronson JE, Liang TP. Decision support systems and intelligent systems. 7th ed. Englewood Cliffs, NJ: Prentice Hall; 2004.

[7] Legendre P, Legendre L. Numerical ecology. Amsterdam: Elsevier; 1998. p. 317–341.

[8] Zar JH. Biostatistical analysis. Englewood Cliffs, NJ: Prentice Hall; 1999. p. 478-481.

[9] Ludwig JA, Reynolds JF. Statistical ecology: a primer of methods and computing. New York: Wiley Press; 1988. p. 337.

[10] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. Los Altos, CA: Morgan Kaufmann; 2005.

[11] Niven PR. Balanced scorecard for government and nonprofit agencies. John Wiley and Sons Inc; 2003.

[12] The European Health Report, Health Systems Performance Assessment Methods, Annex 1, 2005. http://www.euro.who.int/document/e76907.pdf).

6.5 Primary Health Care Network Monitoring: A Hierarchical Resource Allocation Modeling Approach

The provision of useful information to users by monitoring systems can improve their efficiency, security, and satisfaction.

This chapter presents an article on Primary Health Care Network Monitoring, which was published in the International Journal of Health Planning and Management (Pur, et al., 2010). This paper gives a comprehensive description of the proposed methodology for model development and presents its application for the monitoring and assessment of the resource allocation in the PHCN. The paper provides the basic concept of the model as well as the detailed description of the model's indicators, visualizations, and techniques. The model is developed in accordance with the proposed methodology (Chapter 4).

INTERNATIONAL JOURNAL OF HEALTH PLANNING AND MANAGEMENT
Int J Health Plann Mgmt 2010; **25**: 119–135.
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/hpm.1001

Primary health-care network monitoring: a hierarchical resource allocation modeling approach

Aleksander Pur[1]*, Marko Bohanec[2,3],
Nada Lavrač[2,3] and Bojan Cestnik[2,4]

[1] *Ministry of the Interior, Ljubljana, Slovenia*
[2] *Jožef Stefan Institute, Ljubljana, Slovenia*
[3] *University of Nova Gorica, Nova Gorica, Slovenia*
[4] *Temida, d.o.o., Ljubljana, Slovenia*

SUMMARY

Management of a primary health-care network (PHCN) is a difficult task in every country. A suitable monitoring system can provide useful information for PHCN management, especially given a large quantity of health-care data that is produced daily in the network. This paper proposes a methodology for structured development of monitoring systems and a PHCN resource allocation monitoring model based on this methodology. The purpose of the monitoring model is to improve the allocation of health-care resources. The proposed methodology is based on modules that are organized into a hierarchy, where each module monitors a particular aspect of the system. This methodology was used to design a PHCN monitoring model for Slovenia. Specific aspects of the Slovenian PHCN were taken into account such as varying needs of patients from different municipalities, existence of small municipalities having less than 1000 residents, the fact that many patients visit physicians in other municipalities, and that physicians may work at more than one location or organization. The main modules in the model are focused on the overall assessment of the PHCN, monitoring of patients visits to health-care providers (HCPs), physical accessibility of health services, segment of patients in municipalities who have not selected a personal physician, assessment of the availability of HCPs for patients, physicians working on more than one location, and available human resources in the PHCN. Most of the model's components are general and can be adapted for other national health-care systems. Copyright © 2010 John Wiley & Sons, Ltd.

KEY WORDS: primary health-care network; resource allocation; health-care disparities; data analysis; data mining

INTRODUCTION

Imagine being a manager responsible for the organization of a primary health-care network (PHCN). Suppose that you get a call from a local politician who claims that

*Correspondence to: A. Pur, Ministry of the Interior, Štefanova 2, SI-1501 Ljubljana, Slovenia. E-mail: aleksander.pur@policija.si inhabitants of his municipality are health-care deprived, and need more physicians. The problem is how to react in this situation. Will you oppose or support the politician's request, and how will you explain your answer? Do you have sufficient knowledge about the PHCN and inhabitants in that and other municipalities? Do you have sufficient – and sufficiently accurate, accessible, and well-organized – data about the local patients, their needs, and available local health-care services? Are the PHCN requirements, constraints, and management rules sufficiently well-defined? In any case, this is not an easy task.

Management of a PHCN must be based on rational decisions, supported by analyses of actual health-care data. These analyses can be provided by an appropriate monitoring and assessment system that considers a national PHCN planning strategy. This paper presents the approach that was taken to develop such a system in the Republic of Slovenia. On one hand, the approach takes into account the specific characteristics of the Slovenian PHCN, but, on the other hand, proposes a general methodology for structured development of monitoring systems. Both aspects are addressed in this paper.

The Slovenian national health-care system (HCS) is composed of different health-care providers (HCPs), and is divided into the primary, secondary, and tertiary health-care levels. The PHCN consists of four sub-systems: general practice, gynecology, pediatrics, and dentistry. According to the national health-care program (MoH, 2000), municipalities are appointed to manage PHCN on their territory according to the model that must be prescribed by the Slovenian Ministry of Health (MoH).

Recently, the MoH has initiated and financially supported the activities to develop a new, thoroughly revised PHCN monitoring model. Considering the Slovenian National Program of Health (MoH, 2000), the main strategy used in this model is focused on providing good and equal availability and physical accessibility of the PHCN for all residents of Slovenia.

In accordance with these aims, the PHCN monitoring model has to be able to assess the patients physical accessibility to the HCPs, assess the availability of physicians for patients, identify health-care deprived groups of residents, and provide other information about expected and unexpected anomalies related to the inappropriate allocation of PHCN resources. The model has to identify these anomalies as quickly and comprehensibly as possible in order to provide timely alert and prevention from unnecessary bad consequences.

Considering the characteristic of the national HCSs, this is not an easy task. Slovenia has about 2 million inhabitants, and it is divided in 210 municipalities where some of them have less than 1000 residents. Hence, many patients visit HCPs in other municipalities. Another problem is that not all physicians work full time and some of them work on two or three different locations. All these characteristics request a specific approach to the design of the monitoring model.

The developed approach can also be adapted for use in other national HCSs. Most of the modules are general in the sense that all HCSs are composed of similar elements analyzed in this model such as inhabitants living in some characteristic conditions, health-care services needed by these inhabitants, HCPs, and physicians who provide the services. Moreover, a similar approach can be useful even for the monitoring of systems as different as police or supermarket chains.

The design of the proposed PHCN monitoring model is based on a general-purpose methodology that defines monitoring modules and organizes them into a hierarchy. Each module monitors a particular aspect of the PHCN, which is of interest for decision-makers and managers of the system. Modules include all important elements of the PHCN: patients, medical staff, and medical organizations, as well as their geographical and organizational relationships. Typical aspects about PHCN are: assessment of the PHCN considering the allocation of health-care resources, availability of the PHCN resources for patients, visit rate of patients, and the physical accessibility of the PHCN for patients. The modules use the data collected by standard procedures in HCPs, and provide on this basis a number of data analyses and visualizations. The approach is not limited to any particular data analysis methods, but rather provides means to include and combine different and possibly innovative data analysis and data presentation methods.

The model was developed in accordance with the available data related to the PHCN from the following databases:

- Social Security database is the data about HCPs together with assigned patients per individual general practitioner and the patients with social security.
- The database of physicians and dentists, provided by the Slovenian Medical Chamber.
- The database of the National Institute of Public Health containing data about Slovenian health centers.
- The database of the Slovenian Statistics Bureau concerning the demographic and geographic distribution of citizens and municipalities.

This paper is structured as follows. In the next section, we describe some related approaches to assessments of HCS in other countries. The third section describes the basic methodology used in the development of the monitoring model. The fourth section presents the specific structure and modules of the developed PHCN monitoring model for Slovenia. The fifth section presents the SWOT analysis, while the sixth section provides some conclusions and directions for further work.

RELATED WORK

A national HCS is a complex system in which large quantities of data about different processes supported by information systems are produced daily (Babulak, 2006; Haux, 2006). These and other data related to HCS, such as demographic or geographic data, include useful information about health systems. Thus they are periodically or continuously analyzed in order to provide government officials, development managers, and civil society with the information for improving internal processes, planning (Williams, 1999), and resource allocation. The analyses are based on a number of carefully designed financial and non-financial indicators of HCS performance.

In order to provide an HCS assessment framework, the World Health Organisation (WHO) has defined the following goals (WHO, 2000): population health, responsiveness, and fairness in financial contribution. The population health is

DOI: 10.1002/hpm

assessed by two indicators aimed at overall population health and health distribution. The responsiveness indicators address the system performance relative to non-health aspects considering population's expectations of how they should be treated by HCPs. The fairness in financial contribution of population is based on the distribution of households' financial contribution calculated by household survey data.

The Australian National Health Performance Committee (NHPC, 2001) presented another health-care assessment framework, which is adapted from the Canadian Institute for Health Information (CIHI, 1999), and aimed at the overall assessment of HCSs. This framework is implemented in Australian public sector mental health services (ISC, 2005). The indicators cover three main aspects of the HCS: health status and outcomes, determinants of health, and health system performance.

Another HCS assessment framework is the Euro Health Consumer Index (EHCI) (HCP, 2007). The EHCI is composed of 27 indicators that are focused on patient rights and information, waiting time, outcomes, generosity, and pharmaceuticals.

In comparison with the assessment frameworks described above, our PHCN monitoring model is not aimed at the overall assessment of the HCS, but at the assessment of primary health-care resource allocations considering health-care capacity and the patients' needs. Therefore, we have to develop our own indicators.

In order to avoid unmanageable proliferation of indicators, caused by attempting to cover every important aspect of a PHCN (Perera *et al.*, 2007), the indicators in our model are hierarchically organized. In this way they provide both simple top-level and detailed low-level views on the PHCN.

Many frameworks for HCS assessment include indicators composed of sub-indicators. These composed indicators are usually calculated as a weighted sum of the values of sub-indicators. This is not always the best solution. For example, considering the assessment framework proposed by WHO, an HCS with a high population health, low responsiveness and bad fairness in financial contribution has the same overall assessment as the HCS with a low population health, high responsiveness and good fairness. The different HCSs have the same overall assessment. To avoid such problems, the indicators in our PHCN monitoring model are combined using multidimensional graphs, pivot tables, and techniques of multicriteria decision models.

Many indicators, such as the *responsiveness* (WHO, 2000), are based on representative households or key informant surveys using either face-to-face or postal interviews. A major concern with such survey instruments is that people from different cultures or socioeconomic backgrounds usually have different expectations. Consequently, the differences in their answers could reflect the differences in their expectations rather than the variations of HCS responsiveness. In our model, the assessments are based on data already provided by standard procedures in HCS. This makes a considerable difference in the type, quality and quantity of collected input data, affects the selection of indicators, and justifies a different methodological approach taken in our case.

In summary, to address specific characteristics of the national PHCN, available PHCN data and specific objectives of our monitoring systems, we had to design our own set of indicators and methods as described in the following two sections.

DOI: 10.1002/hpm

MODEL DEVELOPMENT METHODOLOGY

Despite many frameworks related to performance and activity monitoring, such as data-driven decision support system (DSS) (Power, 2002), performance monitoring, business performance management (BPM), business activity monitoring (BAM) (Dresner, 2003), etc., there is a lack of methodologies for presenting the overall concept of the monitoring and assessment model composed of different data analysis and data presentation methods.

Our approach to the design of the PHCN monitoring model is based on hierarchically connected *modules* (Figure 1 shows an example of such a hierarchy, which is explained later in PHCN Monitoring Model Section). Each module is aimed at monitoring a particular aspect of the PHCN, which is of interest for decision-makers and managers of the system. Typical aspects about PHCN are: physical accessibility of the PHCN, the availability of the PHCN resources for patients, and visit rate of patients.

Each monitoring module involves a number of *monitoring processes* (MPs), which are gathered according to a particular aspect of the PHCN. Here, the term "monitoring process" is used in a broad sense and denotes a periodically or continuously performed data analysis process without distinguishing between

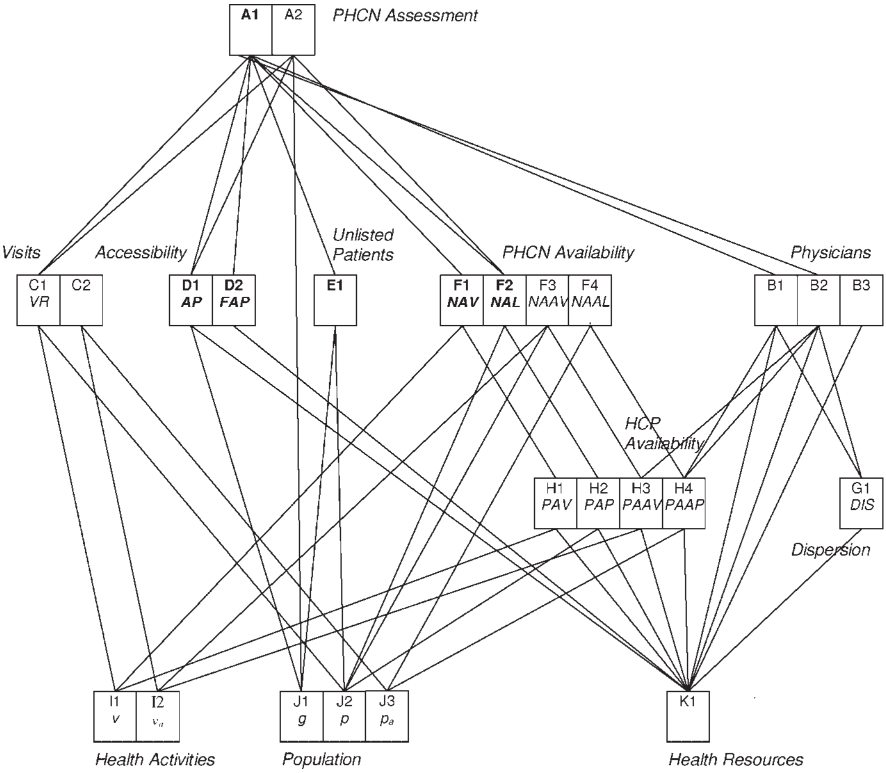


Figure 1. Hierarchical structure of modules and monitoring processes

DOI: 10.1002/hpm

''monitoring,'' ''validation,'' and ''assessment.'' Notice that the term MP has a wider meaning than the term ''indicator,'' because it covers a data analysis process, which may include one or more indicators.

Each MP is characterized by the monitoring objectives, input data, data collection methods, constraints on the data, data dimensions, data analysis methods, output data, the role of output data for PHCN monitoring and planning, the output data presentation methods, target output values, the possible side effects of MPs, evidence base of the previous use of indicator, security requirements, different threats inferred from output data, and the users of these components.

Among these components, the *data analysis methods* transform the *input data* to *output data* represented by some *data presentation methods* according to the given *monitoring objectives*. Usually, the output data consist of different indicators presented by various *data presentation methods*. Again, the term ''data presentation'' is used in its broad sense and includes – mostly visual – presentations of both single data elements and more complex patterns, and it does not explicitly distinguish between ''data,'' ''information,'' ''pattern,'' and ''knowledge.''

The *target* is a level of performance that the organization aims to achieve for a particular activity. Information about *data collection* defines how and how often the data has been collected or needs to be collected. For example, data can be collected by representative surveys or by standard procedures in organizations according to some refreshment rate. The *data constraints* define the valid input and output data. The *security requirements* define the use and management of the MPs and of the data.

This approach is not limited to any particular data presentation or analysis method. In principle, any *data presentation method* can be used such as pivot tables, charts, network graphs, and maps. The same holds for *data analysis methods*, which can include Structured Query Language (SQL) procedures, On Line Analytical Process (OLAP) techniques for interactive knowledge discovery, as well as knowledge discovery in databases (KDD) and data mining methods (Han and Kamber, 2001) for discovering important but previously unknown knowledge.

The *output data* from the MPs can be classified in different ways. In accordance with their role in the decision process, they are classified into two groups. The *output data* in the first group are *formal*: they are based on standards defined by the PHCN management, and formally describe how well the organizational system is achieving its quantifiable objectives. In our model, the MPs providing information from the first group are highlighted in bold (Figure 1). The second group provides *informal* data that describe PHCN activities and performances.

Considering the influence of managers' decisions on a certain aspect of the PHCN covered by MPs, the *output data* are also classified into two groups. The output data in the first group, such as the number of physician per population, can be changed by appropriate decisions of managers, for example, by assigning physicians' positions. The data in the second group, such as the structure of population's age and gender, cannot be directly influenced by managers' decisions. It is included in the monitoring model primarily to show the current state of PHCN.

With respect to organizational goals, the *output data* in the model are classified into *lead* and *lag* (Niven, 2003). The *lead* ones show the performances that have influence on achieving the goals, whereas the *lag* are related to the degree of

DOI: 10.1002/hpm

achieving the goals. For example, the visit rate of patients to PHCN may not be considered as the goal of the PHCN so it is classified as a *lead* metric, but the equal accessibility of PHCN for all patients is usually considered as one of the main goals of this network, so the corresponding metric is classified as *lag*.

The output *data presentations methods* used in our model are classified in accordance with data organization into three categories: text only, tables, and information graphics. *Information graphics* – or ‘‘infographics’’ – are visual presentations of information, data, or knowledge such as graphs, charts, flowcharts, diagrams, and maps. The presentation methods are also divided in accordance with the used techniques into *static* and *dynamic*. Static are data presentations that could be printed on the paper without any loss of data or functionality. In contrast, dynamic presentations provide full functionality only on a computer screen. Thus a typical static graph is presented as a picture without interactive functions, and a typical dynamic graph may be combined with OLAP techniques such as drill-down, slice, and dice (Power, 2002).

The proposed PHCN monitoring model addresses also the possibility of side effects that may be induced as a reaction to the monitoring. Such an unwanted consequence is the so-called ‘‘perverse’’ learning that happens when organizations or individuals learn which aspects of performance are measured and which are not, so they use that information to manipulate their assessments (Meyer and Gupta, 1994). For example, among cardiac surgeons in New York whose individual unadjusted patient death rates have been published regularly, there has been a tendency to avoid taking on high risk cases with a subsequent increase in mortality of patients at risk for cardiac surgery (Dranove *et al*., 2002). The unintended consequences could also be caused by profoundness monitoring of individuals (physicians); the monitoring system that severely limits individual freedom could be counter-productive.

In order to improve the comprehensibility of the model, the modules are hierarchically organized. Therefore, the modules at the top level include the main MPs, which are hierarchically connected with MPs at lower levels in the hierarchy. Each *connection* represents a data channel that connects the outputs of lower-level MPs with the inputs of higher-level MPs. In principle, each connection denotes that the output data of lower-level MPs can help to explain the output data of MPs at higher levels. For example, the MP aimed at the main assessment of PHCN can be connected with the lower-level modules addressing physical accessibility of PHCN, availability of PHCN resources for patients, and the ratio between listed and unlisted patients. This functionality, which enables the users to move from data presented by higher-level MPs to appropriate data presentation of lower-level MPs or *vice versa*, can be provided by a suitable user interface.

PHCN MONITORING MODEL

The aim of the PHCN monitoring model is to improve decisions related to the appropriate allocation of health-care resources to those who need it. According to the methodology described in third section, the model consists of hierarchically connected modules. These modules are shown in Figure 1. Each module includes at least one MP denoted by a letter and number. The letter indicates the module that contains the MP, and the number indicates the index of the MP in the module. Some MPs provide output data in the form of named variables; in this case, abbreviated names of these variables are also shown in the corresponding MP rectangles.

PHCN assessment

The top-level module *PHCN Assessment* addresses the main aspect of the PHCN that helps to identify inequalities and other anomalies related to the allocation of healthcare resources. This aspect presents general information rather than detailed explanation of anomalies; detailed explanations are provided by lower-level modules. Thus, the MPs in the *PHCN Assessment* module are based on analyses of output data from lower-level MPs. In principle, the MPs in this module must be able to process several input parameters using different methods of KDD and techniques of multicriteria decision models.

The module *PHCN Assessment* includes two MPs A1 and A2. The A1 presents the main aspects of the PHCN in the form of short clauses, such as "more (or less) physicians (in terms of FTE, full time equivalent) is needed." The clauses are automatically created on the basis of input data provided by the lower-level MPs and rules that compare input data with the prescribed target values.

MP A2 is another MP in this module, and presents the main aspects of the PHCN in municipalities by a multidimensional chart (Figure 2). Each municipality is represented by a dot, which displays four dimensions:

- The dot's horizontal position shows the visit rates of listed patients from municipalities (VR, see MP C1).

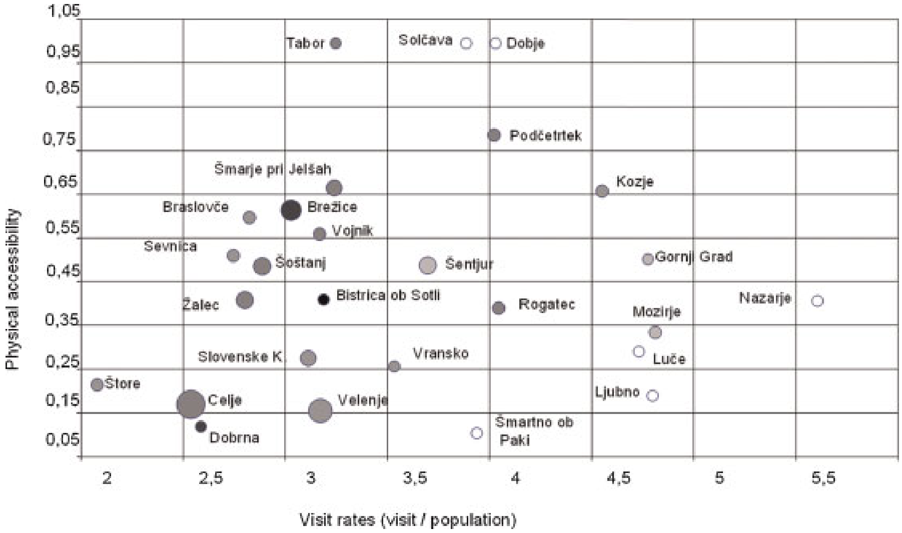


Figure 2. The overall view on the PHCN sub-segment of the general practice for the Celje region (year 2003)

DOI: 10.1002/hpm

- The vertical coordinate corresponds to the physical accessibility of patients to providers (AP, see MP D1).
- The dot color intensity represents the network availability for visits (NAV, see MP F1).
- The dot diameter is proportional to the number of inhabitants in the municipality (g, see MP J1).

In this way, municipalities that have average values of VR and AP appear in the middle of the chart. The outliers represent more or less unusual municipalities with respect to particular aspects of PHCN. A more detailed and accurate explanation of reasons for such municipalities' positions in the chart is provided by lower-level modules B–F (see Figure 1).

Physicians

The module *Physicians* covers the human resources included in the PHCN. This important aspect is covered by MPs B1, B2, and B3 (see Figure 1). A PHCN manager has to know where more (or less) physicians or nurses are needed now or in the future, particularly because the health studies take many years to complete. This information is provided by OLAP based on the MP B1 that presents the aspects of physicians characterized by their specialization, age, gender, workload, location where they work, and *dispersion* (see Figure 1). On the other hand, MP B2 is aimed at discovering interesting relations in physicians' data by an association rules discovery method (Srikant and Agrawal, 1996). In this model these rules show characteristics of physicians in some municipalities.

The managers of the PHCN also needs to know the education level and obtained licenses of physicians, because in the past some employed physicians were discovered not to have a formal qualification for their job. Thus, MP B3 is aimed at the monitoring of physicians' and dentists' qualification for the job they actually performs. The main performance indicator is the physician's specialization degree, granted by the Slovenian Medical Chamber, which must be verified every 7 years. This specialization degree is a prerequisite for getting a license for employment in an area of medicine. To monitor the suitability of physicians for the job they perform we have used a network visualization technique (Batagelj and Mrvar, 2006).

Visits

The module *Visits* addresses the frequency of patient visits to HCPs. Thus, MP C1 shows an average visit rate (VR) of listed patients from a municipality. VR is a ratio between the number of visits of listed patients from municipality and the number of listed patients in that municipality.

MP C1 is further improved by MP C2 that is focused on the ratio between the actual visit rate and the expected visit rate of the same listed patients over the same period of time. For example, in the chart in Figure 3, the vertical axis shows the ratio between actual visit rates and the expected rates. The chart clearly reveals unusually high visit rates of population under 20 to the general practice in some municipalities,

DOI: 10.1002/hpm

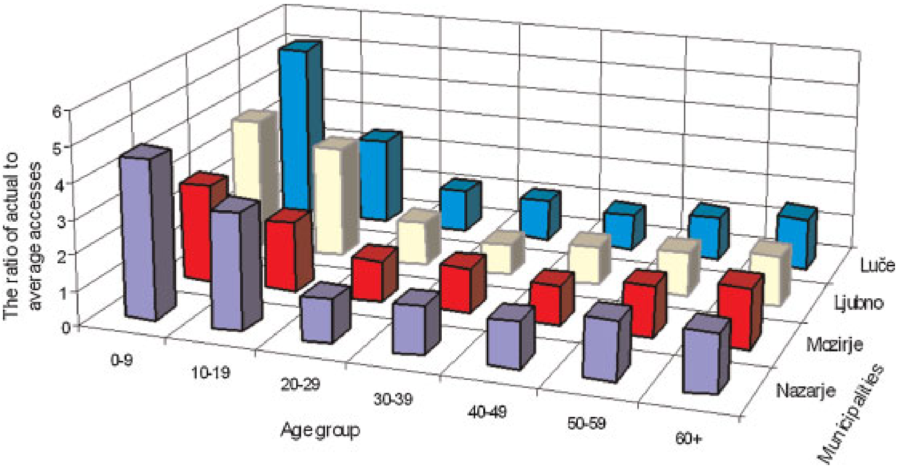


Figure 3. The ratio between actual and expected visits of patients to general practice HCPs (year 2003)

especially in Luče and Nazarje, compared to other municipalities (not shown in this figure, due to space constraints).

In general, the increased visit rate of patients to the HCPs can indicate unhealthy living environments or unhealthy habits of some population groups. Consequently, in addition to PHCN planning, the visit rates are also important for monitoring the health level of population. The MPs in this module are only informal indicators because the patient's visit rates are reflected in results of MPs F1 and F3.

Accessibility

The *Accessibility* module is aimed at monitoring the physical accessibility of patients to HCPs. The module includes two MPs aimed at monitoring the physical accessibility of patients to HCPs. MP D1 (see Figure 1) measures the accessibility for patients (AP) that shows the proportion of the population that needs excessive time to get to the nearest HCP than it is the maximal acceptable travel time, considering the length and the category of roads. Because this proportion depends on the maximal acceptable travel time for the patients to the nearest HCP, these values have to be chosen carefully. In our case, the maximal acceptable travel time was set statistically as one standard deviation above the national average. Generally this time must be prescribed by the responsible authority.

The second MP, D2, provides full accessibility for patients (FAP). It assesses the proportion of population that need more time than the prescribed time to get to the nearest HCP open 24 h daily. The access time computation takes into the account the road length and the road category. The difference between FAP and AP is that first one considers only the HCPs working for 24 h, and the second one considers all the public HCPs. In the model the results are presented similarly by a table and a map (data not shown). Because the accessibility of HCPs is important for patients, the

DOI: 10.1002/hpm

MPs D1 and D2 are used as formal indicators which affect the decisions related to health-care resource allocation.

Unlisted patients

The module *Unlisted Patients* monitors the proportion of unlisted residents in municipalities. The unlisted residents are those without their own physicians. The proportions of unlisted population, provided by MP E1, is calculated as the ratio between the number of unlisted patients and all the population in a municipality. At a primary level the residents are usually assigned to their physicians. This module could show some health-care deprived areas where residents rarely visits the HCPs for different reasons, and usually individuals do not have their own physicians. Because MP E1 provides important information about the residents in municipalities, it is used as a formal indicator.

PHCN availability for patients

The module *PHCN Availability for Patients* is focused on the assessment of the availability of the PHCN for listed patients considering the health-care capacity and needs of the patients. Again, the listed patients are those who have their own physicians. This assessment shows how the PHCN capacity is adapted for the patient needs. The providers in the PHCN with insufficient capacity are overloaded, and their patients are health-care deprived. On the other hand, the providers with excess capacity increase the costs of the PHCN. In general, the assessment of the availability is based on the ratio of the PHCN capacity available for patients to the demand for this capacity from the same patients. Because many patients access HCPs in the neighboring or even more distant municipalities, some municipalities do not have their own HCPs, and some physicians do no work full time, the assessment of PHCN availability for patients in these municipalities is not proportional to the simple ratio between the number of physicians and the number of inhabitants. Thus we have to develop new indicators. This aspect of the PHCN is covered by the next four MPs:

- MP F1 (see Figure 1) provides NAV that is aimed at the assessment of the PHCN availability considering the visits of the listed patients and the PHCN working time. The PHCN working time is the time when the physicians of HCPs included in the network are available for their patients. The NAV is proportional to the average PHCN working time that could be spent for the visits of patients from the municipality.
- MP F3 (see Figure 1) provides network availability for adjusted visits (NAAV). Because the NAV indicator does not consider that different patient treatments require different working time, this assessment is improved by MP F3 that gives NAAV, which is proportional to the ratio between the PHCN adjusted working time spent for visits from the municipality and adjusted visits from the same municipality.
- MP F2 indicator provides the network availability for listed patients (NAL) aimed at measuring the availability of the PHCN for listed patients from a certain area.

DOI: 10.1002/hpm

The NAL patients indicator is defined as the ratio of the PHCN working time available for listed patients from the municipality in a certain period (usually 1 year) and the number of the same listed patients.

- MP F4 considers that different age–gender groups of listed patients spend different amounts of physician's working time for treatments. The output of this MP is called network availability for adjusted listed patients (NAAL), which is proportional to the ratio between available working time of PHCN for listed patients from a municipality and anticipated working time considering age and gender of patients from the same municipality.

The MPs based on patients' visits, such as F1, F3, H1, and H3 (see Figure 1), show the availability of the PHCN in the past. In contrast, the MPs based on the number of listed patients, such as F2, F4, H2, and H4, provide an anticipated availability of the PHCN. This anticipation considers characteristics of listed patients such as gender and age. In general, the assessment of workload for the past shows the anomalies in the PHCN that have already happened, on the other side the anticipated assessment of workload provides the information that could prevent future anomalies in the PHCN, but they are usually less accurate. In the model the output data from these MPs are presented by tables and charts based on OLAP techniques, and maps based on Geographic Information System (GIS) techniques.

Dispersion

The module *Dispersion* addresses the problem that some physicians work on more than one location. Depending on the requirements, a physician may work on more than one location, but this dispersion usually means additional workload for physicians and their lower availability for patients at one of the locations. The dispersion is provided by MP G1.

HCPs availability

The module *HCPs Availability* is focused on the assessment of the HCPs and physicians availability for their patients. This assessment shows how the capacity of each HCPs is adapted for the patient needs. The HCPs with insufficient capacity are overloaded, and their patients are health-care deprived. On the contrary, the excess capacity increases costs of providers. Generally, this assessment is based on the ratio between the capacity of HCP and the needs of patients listed on this provider. The main difference between the MPs in this module and processes in the module *PHCN Availability* (see Figure 1) is that the former processes are aimed at the assessment of the availability of HCPs for patients listed on the provider, and processes in the latter module are aimed at the assessment of the availability of the PHCN for patients from a certain region. The perspective provided by this module is covered by the next MPs:

- MP H1 is aimed at the assessment of the HCP availability considering the patient's visits to HCP and HCP working time. An indicator providers availability for visits

DOI: 10.1002/hpm

(PAV) calculated by this MP is proportional to the average HCP working time that can be spent for a patient's visit.

- MP H3 gives provider's availability for adjusted visits (PAAV) aimed at the availability of an HCP for its patients considering the visits adjusted in accordance with the expected physician's working time needed for treatments. The PAAV is defined as the ratio of the available working time of the HCP for their patients in a certain period to adjusted visits in the same period.
- MP H2 provides the assessment of the providers availability for patients (PAP), which is proportional to the average HCP working time that can be spent for listed patients in a certain time period (usually 1 year). It is defined as the ratio of the HCP working time to number of patients listed to the provider.
- MP H4 consider that different age–gender groups of listed patients require different HCPs workload, where listed patients are adjusted in accordance with the expected physician's working time spent for their treatments. We call this assessment providers' availability for adjusted patients (PAAP), and it is proportional to the ratio between the working time of HCP and expected working time needed for listed patients. The estimation of expected working time is based on the age–gender structure of the listed patients.

The monitoring of physicians' and HCP's availability for patients based on PAP and PAAV may induce unwanted side effects. For example, to improve these indicators, physicians or HCPs may arrange more visits to their patients than necessary in order to show their lower availability for patients and higher workload. This manipulation could be prevented by using additional indicators (PAP and PAAP) that are based on the number of listed patients and not on the number of their visits.

Health activities

The *Activities* module provides the basic data about health activities done by HCPs, which are aimed at improving the health of patients. The actions analyzed are the patient's visits to HCPs. The data about these visits, such as time, HCPs, and diagnosis, are provided by MP H1. MP H2 provides the visits adjusted according with the expected time for health treatment. In H2, the visits are grouped in accordance with the expected time. Each group is weighted considering the physician's working time used for the treatment. Thus the values of adjusted visits is proportional to the physician's time spent for health treatment. The output data from this MPs can be combined with other data or presented by different methods.

Population

The module *Population* is focused on data about the population serviced by the HCPs included in the PHCN. The first MP, J1, provides different data about the population aggregated across municipalities where they live. Most of this data was obtained by the Statistical Office of the Republic of Slovenia. The second MP, J2, provides data about listed patients such as gender, age, and municipalities where they live. Most of

DOI: 10.1002/hpm

this data was obtained from the Institute of Public Health of the Republic of Slovenia. This module also includes MP J3 that provides the expected working time of physicians spent for a listed patient. The listed patients are grouped according to their age and gender, and each group is weighted in accordance with the expected working time spend for all treatments in a certain period. Usually this period of time is 1 year.

Health resources

The module *Health Resources* provides the basic data about physicians and HCPs. This is done by MP I1 that provides the different data about physicians such as age, gender, HCP where they work, FTE, and education. The main sources of this data are the National Institute of Public Health, Health Care Institute Celje, Slovenian Social Security Database, and Slovenian Medical Chamber.

SWOT ANALYSIS

The presented PHCN monitoring model provides the main aspects and detailed information about the network from the patient's point of view. The strength of the model is based on hierarchically connected modules. Each module, composed of at least one MP, is aimed at a particular aspect of the PHCN (Figure 1). The MPs in the top-level module provide the main picture of the PHCN considering the allocation of health-care resources, and the MPs in lower-level modules provide the detailed information abut particular aspects of the network. This hierarchically structured model is appropriate for top-down monitoring.

For example, top level MP A1 (Figure 1) indicates that patients in municipality Nazarje have low value of NAV (see MP F1). The multidimensional chart (Figure 2) provided by MP A2 implies that high visit rate (VR) (see MP C2) in Nazarje is the reason for a low value of NAV in Nazarje. The three-dimensional bar chart (Figure 3) provided by MP C2 shows detailed information about the VR in Nazarje. This chart shows that patients under 20 from Nazarje have more than three times higher VR than expected, considering the average VR for these age groups in the Celje region. This example, based on real data, shows how the model indicates the anomalies in the PHCN. The model is also able to reveal possible reasons for these anomalies (increased visit rates) and detailed information about these reasons (increased visit rates of patients under 20). Considering the information provided by the model, the decisions related to the PHCN can be justified by logical arguments based on the available data.

Another strong point is that the model utilizes different data presentations and analysis methods. For example, data is analyzed by a broad collection of methods, including SQL procedures, OLAP techniques (see MP C1), GIS techniques (MPs D1 and D2), rule-based models (MP A1), and data mining methods such as association rules discovery (MP C2). The results are presented by automatically generated text (MP A1), different charts (MPs A2, B1, and C1), visualization of association rules (MP C2), network graphs (MP B3), and maps (MPs D1, G1, and F1).

DOI: 10.1002/hpm

The size of districts monitored by the PHCN monitoring model is not limited. The model can provide useful information about municipalities with less than 1000 residents and even those without HCPs. This is possible because the methods for PHCN assessments (see the module *PHCN Availability*) consider the migration of patients from and into other municipalities.

The weaknesses of our approach to PHCN monitoring are manifested in potentially misleading information provided by the model and unintended consequences of the monitoring. Despite many interesting methods used in our PHCN, the threat that misleading data can lead to wrong decisions, and consequently to negative impacts on the PHCN, cannot be completely avoided. The main reasons for the misleading information are deficient and wrong input data, incorrectly selected methods of data analysis, and inappropriate interpretation of the results. We also have to consider the unintended consequences of the PHCN monitoring such as ''perverse'' learning (Meyer and Gupta, 1994) and those caused by too profound monitoring of individuals (physicians).

The opportunities provided by the chosen model design lies in the fact that various aspects of the PHCN provided by the model could reduce some unintended consequences of monitoring. For example, the NAV indicator (see MP F1) shows the availability of the PHCN for patient's visits. If the HCPs availability or workload were assessed only from this viewpoint, it could create a ''perverse'' motivation to increase the number of illness treatments rather to work on their prevention. This can be alleviated by the inclusion of several indicators that, when used together, can indicate such anomalies. In our case, we thus assess HCPs availability using a second indicator, *availability for the listed patients* (see MP F2). Further opportunities are indicated in the Conclusions Section.

CONCLUSIONS AND FURTHER WORK

A large amount of available electronic data from different fields, and many methods for their analysis used by efficient software and hardware, present an opportunity to improve efficiency of monitoring systems. In order to utilize these opportunities, it is important that the monitoring model is not limited to any particular form of input data, data analysis, and data presentation method.

The same approach that we used to design the PHCN monitoring model can be used in other fields. For example, in another project we use this approach to describe a model for monitoring and assessment of employees' activities in order to improve information security in a public organization. In this case, the top level module shows unusual and potentially suspicions associations between employees and other persons or objects. The middle-level modules are aimed at detailed aspects of these associations, and the lowest modules describe the basic data provided by different log files and documents.

In principle, the basic approach of the presented PHCN monitoring model can be used for the monitoring of systems as different as police or supermarket chains. For example, police systems are composed of similar elements, such as inhabitants that need security services and police units that provide these services, as those in the

DOI: 10.1002/hpm

PHCN consisting of patients and HCPs. Each police station is responsible for the security in its region, like the HCP is responsible for the health of patients in its region. By such analogy, different organizations can be actually monitored by similarly designed models.

In further work, we plan to adapt the proposed method to other systems (e.g., a police network). Moreover, our present work in the monitoring of the Slovenian PHCN is not concluded. The proposed model and the presented indicators will become even more important once we start addressing the PHCN quality and the actual populations health level, which is planned in future work.

ACKNOWLEDGEMENTS

We gratefully acknowledge the financial support of the Slovenian Ministry of Education, Science, and Sport, and the Slovenian Ministry of Health which is the end-user of the results of this project. We also acknowledge the National Institute of Public Health experts Rade Pribakovič and Tit Albreht, Health Care Institute Celje expert Tadeja Kopač, as well as Andrej Kobler and Anton Gradišek who have contributed to the results of this study.

REFERENCES

Babulak E. 2006. Quality of service provision assessment in the healthcare information and telecommunications infrastructures. *Int J Med Inform* **75**: 246–252.

Batagelj V, Mrvar A. 2006. *Program for Analysis and Visualization of Large Networks*, Reference Manual. University of Ljubljana: Ljubljana.

CIHI. 1999. *National Consensus Conference on Population Health Indicators: Final Report*. Canadian Institute for Health Information, ISBN 1-895581-56-7.

Dranove D, Kessler D, McClellan M, Satterthwaite M. 2002. *Is More Information Better: The Effects of 'Report Cards' on Health Care Providers*, Working Paper w8697. National Bureau of Economic Research: Cambridge.

Dresner H. 2003. *Business Activity Monitoring: BAM Architecture*, Gartner Symposium ITXPO. Cannes: France.

Han J, Kamber M. 2001. *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers: San Francisco.

Haux R. 2006. Health information systems – past, present, future. *Int J Med Inform* **75**: 268–281.

HCP. 2007. Euro Health Consumer Index. Health Consumer Powerhouse. The EHCI report 2007.

ISC. 2005. *Key Performance Indicators for Australian Public Mental Health Services*, Information Strategy Committee Performance Indicator Drafting Group. ISC Discussion Paper No. 6. Australian Government Department of Health and Ageing: Canberra.

Meyer MW, Gupta V. 1994. The performance paradox. *Res Organ Behav* **16**: 309–369.

MoH. 2000. *Slovenian National Program of Health – Health for All to 2004*. Ministry of Health of the Republic of Slovenia. Official Gazette RS 49/2000.

NHPC. 2001. *National Health Performance Framework Report*. Queensland Health: Brisbane.

Niven R. 2003. *Balanced Scorecard for Government and Nonprofit Agencies*. John Wiley and Sons Inc.: Hoboken, NJ.

DOI: 10.1002/hpm

Perera R, Dowell T, Crampton T, Kearns R. 2007. Panning for gold: an evidence-based tool for assessment of performance indicators in primary health care. *Health Policy* **80**: 314–327.

Power JD. 2002. *Decision Support Systems: Concepts and Resources for Managers*. Quorum Books Division Greenwood Publishing: Westport, CT.

Srikant R, Agrawal R. 1996. *Mining Quantitative Association Rules in Large Relation Tables*. IBM Almaden Research Center: San Jose.

WHO. 2000. *How Well Do Health System Perform?* World Health Report 2000: Health Systems Improving Performance, http://www.who.int/whr/2000/en/whr00_en.pdf, accessed January 26, 2008.

Williams JG. 1999. The use of clinical information to help develop new services in a district general hospital. *Int J Med Inform* **56**: 151–159.

DOI: 10.1002/hpm

7 Advanced Data Analysis Techniques in Monitoring Systems: Summary, Lessons Learned and Recommendations

In accordance with proposed methodology for developing HAMM, the data analyses techniques for monitoring systems are divided into simple indicators and advanced data analysis techniques. The first continuously measures performances and events through indicators such as the accessibility of the HCP (Section 5.4), the availability of physicians for patients (Section 5.8), and visit rate of patients to HCP (Section 5.3). These indicators are calculated by basic data operations such as counting, summarizing, multiplying, dividing, and joining of various data.

On the other hand, the advanced data analyses techniques assess the meaning of the measured values for the past, present, and future of the monitored system. For example, they can assess the extent to which the monitored values are:

- critical - assessment models,
- unusual - outlier detection, e.g., detecting atypical health care resources,
- related between each other - association rules, and
- important for the future - prediction models.

In this dissertation, the advanced data analysis techniques are the assessment models, data mining techniques, and advanced data visualizations. The basic information about these techniques is given in Chapter 3. In accordance with the proposed methodology, these techniques used in the MM have to meet the following criteria:

- They have to be able to identify all relevant expected or unexpected information that can be hidden in a large amount of data. For example, some data mining techniques such as association rules discovery and clustering methods can discover unexpected and useful information about the monitored system (Section 5.2).
- Relevant information about the monitored system has to be provided quickly and in an automated or semi-automated way. Considering the huge amount of daily-generated data in organizations that have to be periodically analyzed, the complicated manual data analyses are unsuitable and cannot be used.
- The advanced techniques have to provide useful information in a clear, comprehensible, and unambiguous way because the operators, supervisors, and decision-makers that are using the monitoring systems are not always skilled in data analysis methods.

The following sections provide some suggestions based on our experience for the implementation of these techniques in the monitoring systems.

7.1 Assessment Models

Monitoring models can include various assessment models that join input parameters and provide new aspects of monitored systems. For example, the models can:

- Assess performances of the monitored system, such as the overall state of the PHCN provided by MP A1 (Section 5.1), recidivism of sexual offenders (Pur, 2001), and various risks.
- Predict future states of the monitored system, such as health of the population, retirement of physicians (Section 6.2) and biological activities (Ladner, et al., 2009).
- Detect outliers (Section 7.2.2).
- Identify discrepancies between expected and current values, e.g., an abnormal number of patients compared to the expected number (Section 6.1).

The proposed PHCN monitoring system includes a rule-based assessment model that identifies important outliers and anomalies of the PHCN, such as the providers, where more (or fewer) physicians are needed. The model is included in monitoring process (MP) A1 (Section 5.1).

In our prior work, we also created a rule-based assessment model that assesses the likelihood that a specific offender will commit subsequent sex crimes (Pur, 2001). The model was created with the use of a tool for the development of rule-based multi-criteria decision models DEXi (Bohanec, 2011).

The assessment models developed by supervised learning techniques join input parameters by rules (e.g., classification, association, clustering, and sequence) or formulas (e.g., regression). The rules can be implemented by Predictive Model Markup Language (PMML) (Pechter, 2009) in the assessment models. The models can be integrated in various analytical systems such as IBM Info Sphere Warehouse (Ballard, et al., 2007). The models are able to manage one, or a few, records in almost real time.

7.2 Data Mining Techniques

7.2.1 Association Rules Discovery Techniques

The association rules discovery technique is described in Section 3.2.1. Usually, this technique produces a large number of rules that are not easy to understand. In order to simplify the explanation and make these rules usable for monitoring systems, we look at them from two perspectives. Both of them are covered by the framework of descriptive rule discovery (Kralj Novak, 2009). The first perspective reveals interesting predefined associations, and the second perspective provides characteristics of head items.

The first perspective considers the fact that some associations are unusual or not allowed. This aspect discovers interesting and unexpected associations between body (*A*) and head (*B*). For example, any association between areas where people live and any illnesses of these people can be interesting and could mean that they live in unhealthy conditions (Pur and Bohanec, 2003).

In our prior research, we assumed that any relation between weather conditions and suicide rate is interesting (Pur and Bohanec, 2003). Using this approach, we found interesting relations, e.g., influence of south winds on the suicide rate in Ljubljana.

A similar example is our analysis aimed at telephone traffic data (Pur and Belič, 2004). We assumed that any relations between telephone calls on days on which individuals are absent from work and called people could be interesting because they show particular relations among these people.

On the one hand, these interesting relations can be implemented in the rule-based assessment models. On the other hand, the association rules discovery technique can be used in the MM to discover new unpredicted and interesting relations.

From another perspective, association rules show the characteristics of head items. The formula for *Lift* (Formula 3) shows how the *Confidence* is greater than expected. If we rearrange the formula (Formula 4), then *Lift* shows how items in body (*A*) characterize the item in head (*B*). In this case, *Lift* is defined as the ratio of the share of characteristic events (*A*) in the area (*B*) to the share of these events in all monitored areas.

$$Lift = \frac{|A \cap B| / |A|}{|B| / |S|} = \frac{|A \cap B| / |B|}{|A| / |S|} \tag{4}$$

In the proposed model for monitoring of the human resources in the Slovenian HCN (Section 6.2), we used association rules discovery techniques to get the characteristics of physicians in municipalities considering their specialization, gender, age, and workload. This advanced data analysis technique provides additional and unexpected information about the monitored system.

7.2.2 Outlier Detection by Clustering

This section describes the use of outlier detection techniques in monitoring models. The approach to detection is based on clustering techniques. In accordance with Hawkins (1980), an outlier is an observation that deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism. Barnett and Lewis (1994) defined an outlier as an observation (or subset of observations) that appears to be inconsistent with the remainder of that set of data.

Many techniques employed for detecting outliers are fundamentally identical but with different names chosen by the authors. For example, authors describe their various approaches as outlier detection, novelty detection, anomaly detection, noise detection, deviation detection, or exception mining. In this paper, we have chosen to call the technique outlier detection (Chandola, et al., 2009).

Outlier detection has a strong application background in fields such as telecommunications, financial fraud detection, medicine, and data cleaning. For example, fraud detection can be based on the monitoring of the purchasing behavior of a credit card owner that becomes unusual when the card is stolen or abused. This technology can also detect fraudulent applications for credit cards and fraudulent usage of mobile phones by monitoring their activities. In medicine, unusual symptoms or test results may indicate potential health problems for a patient, or in public health, the occurrence of a particular disease, e.g., tetanus, scattered across various hospitals of a city may indicate problems with the corresponding vaccination program in that city. Thus, outliers themselves draw much attention, and outlier detection is studied intensively by the data mining community (Breunig, et al., 2000; Ramaswamy, et al., 2000). There are three fundamental types of approaches to the outlier detection based on data analysis (Hodge and Austin, 2004), as presented below.

Type 1: Determine the outliers with no prior knowledge of the data

This approach processes the data as a static distribution, pinpoints the most remote points, and flags them as potential outliers. The Type 1 approach assumes that errors or faults are separated from the "normal" data and thus they appear as outliers. Considering this approach, the outliers can be defined as clusters with one or a few members. Small clusters are often eliminated, since they mean noise but for the outlier detections, they can identify outliers that are exceptional when compared with larger clusters.

Type 2: Model both normality and abnormality

This approach is analogous to supervised classification and requires pre-labeled data, tagged as normal or outlier. The models generated by supervised techniques provide fast identification of outliers in large amounts of data. On the other hand, the learning by examples, where the data has to be tagged, is not always the best solution.

For example, the gathering of abnormal data from aircraft accidents or accomplished terrorist attacks could be extremely costly. The analyses of abnormal data about events such as smuggling of weapons of mass destruction are relatively rare and thus cannot provide useful patterns. Moreover, the smugglers and terrorists can change methods of operating for each action.

Type 3: Model only normality or in a very few cases model abnormality

The authors generally name this technique novelty detection or novelty recognition (Fawcett and Provost, 1999; Japkowicz, 1995). It is analogous to semi-supervised recognition. It requires no abnormal data for training, which is often difficult or expensive to obtain. The approach needs pre-classified data but only learning data marked normal, and it can learn the model incrementally as new data arrives (Pokrajac, et al., 2007). Thus, it is suitable for static or dynamic data. The approach aims to define a boundary of normality in order to recognize any new data as normal if it lies within the boundary (Hodge and Austin, 2004).

Generally, the third approach is the most suitable for outlier detection in monitoring systems because it is sufficiently fast to handle a large volume of dynamic data, and it does not require pre-labeled data, as the Type 2 approach does. On the other hand, when we are not analyzing a large amount of dynamic data, we can use the Type 1 approach based on data clustering. The advantage of the Type 1 and 3 approaches is that we do not need pre-classified data.

In our unpublished work, we tested a Type 1 system for outlier detection. The system was intended at detecting the usual and unusual behavior of undercover agents. To do this, each week was described with attributes such as the number of contacts with suspicious persons, amount of gathered information, spending money, and working hours. Considering these attributes, the weeks were clustered. The clusters with many members identify different but usual behaviors of the agents. The clusters with one or few weeks identify outliers – unusual behaviors of the agents.

The clustering technique used in this system is based on the Demographic Clustering method, which is implemented in InfoSphere Warehose 9.5 (Ballard, et al., 2007). The method can handle numeric and nominal data. The Demographic Clustering kernel implements a non-hierarchical, iterative, and distribution-based clustering algorithm whose similarity criterion is based on the so-called Condorcet criterion. The algorithm tries to find the partition with the maximum Condorcet value (Ballard, et al., 2007).

We identified some interesting outliers such as weeks characterized by low activities and useful information but a large amount of spending money, and weeks characterized by a large amount of useful information and low activities.

We also found clusters that reflect usual behavior of the undercover agents. For example, we found clusters with many members that show usual weeks with low activities, results, and spending money. The information about the agent behaviors can be interesting for their supervisors.

7.2.3 Prediction Data Mining Techniques

The Prediction DM techniques briefly described in Section 3.2.3 can be used in the monitoring systems to generate prediction models that are implemented in the MM (Section 7.1).

We utilized this approach in the model for the estimation of the number patients that can be handled by a community health center (the center's capacity estimated by the number of employed staff) (Section 6.1). The model is based on linear regression techniques.

We also utilized the regression technique in the research project MediNet to create a model for the prediction of the number of patients' visits to health care providers in certain areas (MediNet, 2005).

7.3 Advanced Visualization Techniques in Monitoring Systems

The proposed methodology to design of the monitoring models applies advanced visualization techniques that highlight the results from various perspectives. These perspectives have to be shown in a clear, comprehensible, and unambiguous way because not all users are highly skilled in statistical and data analysis methods. The techniques that we tested and implemented in the monitoring models (Chapter 5, Chapter 6) are presented below.

Geographic map with pie charts

Geographic presentations of icons show the characteristics of monitored areas as well as the adjacent areas and the size of areas. The colors of areas, icons, or graphs on the areas can show many attributes of these areas. For example, the map in Figure 6 (Section 6.2) shows the number and share of patients affected by the retirement of physicians for five years.

The ratio between actual and expected values presented by 3D graph

In order to show the aspects of actual and expected performance values, we can present a ratio between these values by a 3D chart. For example, the ratio between actual and expected visit rates of patients to the HCPs for age groups in municipalities is presented in Section 5.3. There, the expected visit rates are averaged rates for a certain age group.

Scatter graph for outlier detection

In general, a scatter graph can present various dimensions, because each data point can be characterized with position, shapes, size, color, or animation. Moreover, the shapes such as "Chernoff faces" can show additional dimensions. In a scatter graph, the outlier can be detected as a rear data point considering the values of other data points.

For example, the municipalities in a scatter graph (Section 5.1) are presented by dots characterized by size and color that is proportional to the availability of health care providers. This presentation clearly shows the outliers such as municipalities with a low availability of health care provide.

The scatter graph based on OLAP technology can provide additional aspects with interaction techniques such as slice and dice. For example, the scatter plot in Section 5.2 presents the aspects of physicians in municipalities characterized by their specialization, age, gender, workload, and dispersion. That presentation makes exploration in the data even more intuitive.

Scatter graph with regression line

Another view of the disparity between actual and expected values can be provided by a scatter graph with a regression line showing expected values. For example, the regression line on Figure 2 (Section 6.1)5.1 represents the expected number of patients that can be handled by a community health center (the center's capacity estimated by the number of employed staff), with respect to the actual number of patients handled by the center. Health centers that lie under the regression line have an insufficient number of staff compared to the number of actual patients.

Associations presented by graphic tables

Aggregated results can hide great variations. The visualizations such as graphical tables can reveal outliers inside the aggregated values.

For example, the graphic table on Figure 3 (Section 6.3) shows associations between age-gender grouped patients from municipalities and patients grouped by visits to medical specialists. The size of each rectangle in the chart is proportional to the parameter *Support*, and the color of the rectangle depends on the parameter *Lift* (Srikant and Agrawal, 1996). The red rectangles represent strong relations (with *Lift* higher than 4), and the blue ones represent associations where the *Lift* is smaller than one. The parameters *Support* and *Lift* are provided by association rules discovery methods. These graphic tables can also include interaction techniques, such as a detailed explanation of each association in a message box, which is obtained by moving the screen pointer on the selected rectangle (Section 6.3).

Graph-based visualization of parallel dimensions

These visualizations show relations between attributes, which describe entities. The technique clearly shows well the typical thick lines (denoting a high number of cases) and atypical thin lines (low number of cases). Thus, it enables abnormality detection and further analysis of individual discovered anomalies. For example, the visualization of physicians' qualifications for the job they actually perform, which is based on this technique, is included in the proposed health care monitoring model (Section 6.3). Each physician is described by his or her specialization degree, license, and type of patients. The visualization clearly shows typical and atypical relations between these three properties.

7.4 Advanced Data Analysis Tools

This section describes the main tools that were used in monitoring models and processes described in this dissertation. Considering that data analysis techniques vary from simple queries to advanced data mining and visualization techniques, various data analysis tools were used.

We used Microsoft Excel and Access for simple Extract, Transform, and Load (ETL) processes. These tools were suitable for relatively simple and ad-hoc analyses. However, when we have to manage repeatable monitoring processes that include advanced data analysis techniques, these tools are insufficient.

For more advanced data preparation and data analysis tasks, we applied IBM InfoSphere Warehouse (Ballard, et al., 2007). This is an Eclipse-based commercial data warehouse system that provides simple query techniques as well as advanced analytics, such as OLAP, data mining and text analytics. IBM InfoSphere Warehouse can analyze data from various data sources such as DB2, JDBC, ODBC, and ORACLE. It is integrated with data visualization systems Alpha Blocks, and Cognos.

For data visualizations, we used a combination of tools, including MS Excel charts supported with macro code, Pajek, Cognos, Alpha Blocks, and Geographic Information System (GIS) Manifold.

Pajek is a Windows program for the analysis and visualization of large networks (Batagelj and Mrvar, 2006). Manifold System is a GIS that provides desktop application, objects library for programmers, and Internet Map Server (IMS) for web applications.

Assessment models were built by DEXi (Bohanec, 2011), a computer program for multi-attribute decision making. With DEXi, we interactively developed qualitative multi-attribute decision models, which were used to evaluate and assess decision alternatives. For the development of models from data, we used supervised learning techniques as implemented in InfoSphere Warehouse. The models built by supervised learning techniques were integrated in a scorer operator in InfoSphere Warehouse (Ballard, et al., 2007).

According to the definition in Section 1.1, the term monitoring denotes the process of continuously gathering and performing real-time analyses of data about the monitored system. This raises an important issue of repeatability: it is important that the monitoring system can automatically repeat the monitoring processes. Therefore, the monitoring system has to be stable, reliable, and well-documented. The use of commercial tools, such as InfoSphere Warehouse, provides a clear advantage, as the tool integrates many different methods (from data manipulation to data analysis and visualization) and is able to create and deploy an integrated application package.

We also have to stress that the proposed methodology for developing monitoring systems is not focused on technical development of monitoring systems. The methodology supports the development of new basic concepts of monitoring models as well as reusing existing HAMM and GT, but is in principle independent on – and thus flexible with regard to – the applied technology and software tools.

On the other hand, a higher formalization of the proposed methodology can improve development and deployment of the monitoring models. One way towards achieving this formalization is to introduce a new formal terminological framework for GT and MM described by an ontology (Džeroski, et al., 2010). For example, the GT for monitoring and assessment of public service availability and resource allocation (Section 4.2.1) can be used in various fields. This GT is composed of the seven modules that reveal different aspects of the monitored systems and each module is composed of MPs. Moreover, the aspects revealed by modules and included MPs are similar, even if MMs are aimed at different monitored systems. An ontology referring to this GT can formally describe all its modules, all MPs that can be included in modules, all characteristics of the MPs (e.g., input data, results, and techniques), as well as relations between them. Thus, this ontology can provide a clear and formal description of the monitoring system.

Nevertheless, the formally described models are not sufficient for real improvement of the proposed methodology. In order to improve the development and deployment of the formally described MMs, a new development platform is required that would include and support all techniques used in MM. The inductive databases and constraint-based DM techniques (Džeroski, et al., 2010) are an emerging research area that can provide these platforms. The inductive databases include data, patterns (e.g., itemsets, substrings, and subgraphs), and models (e.g. classification trees, regression trees, and Bayesian networks). They are manipulated by a query language that is an extension of a database query.

8 Results and Conclusions

Basically, we cannot imagine the management of any system without its monitoring. Fortunately, the huge amount of available electronic data and many effective techniques for their analyses seem a good opportunity for monitoring various systems. The problem is that the systems combining all these monitoring possibilities can be huge and complicated. For purposes of resolving this problem, we proposed a new methodology for developing complex monitoring models. This methodology was primarily used for the design of the Slovenian PHCN monitoring model.

The proposed methodology is based on the assumptions that the monitored system can be observed from various aspects that are hierarchically connected, and each aspect can be monitored by various data analysis techniques. Furthermore, we presumed that some monitoring concepts are generic. For example, there are different public services, such as police, fire brigade and health care, which share common monitoring elements and processes.

The main strength of the proposed methodology is in the comprehensible modeling of basic concepts as well as the details of the monitoring model, which is possible by the HAMM (Section 4.1). The strength of the proposed methodology is also in reusable monitoring models – the GTs (Section 4.2). We learned that the basic concept of the PHCN monitoring model can be reused for the monitoring of systems as different as police or supermarket chains (Section 4.2.1). By such analogy, the monitoring models in different organizations can be based on the same GT.

The next strength of the proposed methodology for developing MM is scalability. It can be used for small and simple models as well as large and complex models. Moreover, the models can be enlarged and joined. The proposed methodology is also not limited to any data analysis method. In particular, this methodology is appropriate for, but not limited to, implementation of the advanced data analysis techniques such as the assessment models, the DM techniques, and the advanced visualization techniques (Chapter 7). This dissertation provides an extensive collection of suggestions and examples of using these techniques in monitoring models.

8.1 Results Summary

This section presents and discusses the achieved results and discoveries. These are described in the following order: the proposed methodology for designing monitoring models, the implementation and use of the advanced data analysis techniques in monitoring models, and the PHCN monitoring and assessment model used on real data.

8.1.1 Methodology for Developing Hierarchical Assessment and Monitoring Models

The proposed methodology for developing monitoring models is based on HAMM, GT, and advanced data analysis techniques.

The HAMM is a model of a monitoring system that improves its design and comprehensibility. Section 4.1 shows the functionality of HAMM, as well as the phases for the design of the HAMM. The main functionalities of the HAMM are:

- HAMM can comprehensibly describe the basic monitoring concept as well as details of the monitoring system.
- HAMM is not limited to any data analysis techniques.

We learned that the development of the monitoring models can be simplified by reusing existing monitoring concepts. These concepts are described by GT (Section 4.2). The main functionalities of GT are:

- GTs provide the basic templates for monitoring models.
- These templates can be reused for monitoring similar systems in various fields. The examples of GTs are described in the dissertation (Sections 4.2.1 –4.2.3). The first GT was used as a basis to develop the PHCN monitoring and assessment model (Chapter 5).

In the dissertation, we proposed the following phases of the HAMM design (Section 4.1):

1. providing prerequisites,
2. defining relevant aspects,
3. defining indicators,
4. defining MPs and the advanced data analysis techniques, and
5. validation and testing.

8.1.2 Implementation and Use of Advanced Data Analysis Techniques in Monitoring Systems

We proposed the use of advanced data analysis techniques in monitoring systems. Thus, we classified them into three groups (Chapter 3)3 :

- assessment models,
- data mining techniques, and
- advanced data visualization techniques.

Based on our experience in developing various monitoring systems, we provided suggestions for effective use of different data analysis techniques.

Assessment models applied in the following application areas:

- assessing the overall state of the PHCN (Section 5.1),
- assessing the capacity of public services (Section 6.1), and
- predicting the needs for physicians in the future (Section 6.2).

Data mining techniques were used in the following areas:

- finding interesting or not allowed associations (Pur and Bohanec, 2003),
- finding characteristics of certain areas (Section 5.2),
- identifying outliers (Section 7.2.2), and
- creating assessment models (Section 6.1).

We used the advanced data visualization techniques in the following areas:

- showing characteristics of certain areas by multidimensional graph (Section 5.1),
- presenting the qualifications of physicians for their job (Section 5.2),
- identifying outliers (Section 7.2.2), and
- visualizing the digital maps (Sections 6.4).

8.1.3 The PHCN Monitoring and Assessment Model

We developed the model for the monitoring of resource allocations in the Slovenian PHCN (Section 5). The model was developed in accordance with the proposed methodology. It is characterized by hierarchical structure of the crucial aspects of the PHCN, advanced data analysis techniques, and novel indicators. The PHCN monitoring model is based on the GT for the monitoring of the availability and resource allocation of public services.

The model is able to identify health care-deprived groups and areas, populations with poor health condition, increased expenses of HCN, and physicians without license or formal education, and provides other information about expected and unexpected anomalies related to the PHCN.

The model includes some advanced data analysis techniques, such as the assessment models, the DM techniques, and the advanced data visualizations.

Considering the characteristics of the Slovenian PHCN, we proposed new indicators:

- availability of the PHCN for patients (Section 5.6),
- availability of the PHCN for visits (Section 5.6),
- dispersion (Section 5.7), and
- unlisted patients (Section 5.5).

The model was developed and used within the MediMap (2004), MediNet (2005), MediNet+ (2006), MediNet++ (2008) projects. Thus, the methods and indicators included in the proposed model were used on real data, and approved by the Slovenian MoH in the years from 2004 to 2008. The use of the model revealed considerable inequalities in availability and accessibility of the PHCN for population from some municipalities. Iconsistencies regarding physicians' specializations and licenses were observed. As a contribution to PHCN management, this model made an assessment of the number of physicians in municipalities that would have to be replaced in the next five years. The model identified age-gender-grouped patients from certain municipalities that visit medical specialists more frequently than expected. The main results are described in project reports (MediMap, 2004; MediNet, 2005; MediNet+, 2006; MediNet++, 2008).

8.2 Conclusions

We successfully utilized the proposed methodology for designing MM aimed at allocation of the Slovenian PHCN resources (Section 5) as well as other monitoring models (Chapter 6). These applications indicate that such models can improve planning, measuring performances, and indicating anomalies in the PHCN.

One of the weaknesses of the proposed methodology is that a visualized MM composed of many monitoring processes can be huge and complicated, for example the PHCN monitoring model (Figure 6). However, this cannot be avoided, considering that monitoring systems in organizations can include more than one hundred reports.

A big opportunity for the improvement of existing and development new MMs is in the fact that every day more and more data about various systems are stored in electronic forms, and that techniques for their analyses are continuously improved. Considering the characteristics of the world today, such as the dynamic business environment, globalization, and economic and security crises, the improvement of monitoring systems is a necessity.

The main threats of the monitoring systems are manifested in misleading information provided by the model that can lead to wrong decisions. The main reasons for the misleading information are deficient and wrong input data, incorrect methods of data analyses, and wrong interpretation of the results. Other possible threats originate in unintended consequences of the monitoring such as "perverse" learning and those caused by too profound monitoring of individuals (Section 4.1). We have to be aware that relying only on the monitoring systems can lead to wrong decisions. After all, full responsibility for decisions should remain on the decision makers not on the information system.

For further work, we will use the proposed methodology to develop a monitoring model aimed at the monitoring of users' current activities in order to provide additional information related to these activities. Thus, the model will automatically analyze data about user's current activities, such as reading a large amount of electronic messages or electronic document, and assess which additional information can help the user to manage this large amount of information.

9 Acknowledgement

I want to thank many people for a great variety of reasons. Firstly, I wish to thank my supervisor Marko Bohanec, who—besides being my supervisor—was also my mentor, adviser, motivator, and friend. I also want to thank Nada Lavrač and Bojan Cestnik for invaluable motivation and recommendations.

Let me express my respect to my colleagues at the MediMap (2004), MediNet (2005), MediNet+ (2006) and MediNet++ (2008) projects from the Department of Knowledge Technologies at Jožef Stefan Institute, Ljubljana, Slovenia.

I also acknowledge the National Institute of Public Health experts Rade Pribakovič and Tit Albreht, Health Care Institute Celje expert Tadeja Kopač, as well as Andrej Kobler and Anton Gradišek who have contributed to the design of the PHCN monitoring and assessment model.

The project MediNet was financially supported by the Health Care Institute Celje, and the three MediNet projects were funded by the Ministry of Health of the Republic of Slovenia.

10 References

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A. I. (1996). Fast Discovery of Association Rules. Advances in Knowledge Discovery and Data Mining, pages 307–328.

Astel, A., Astel, K., Biziuk, M., Namieśnik, J. (2006). Clasification of Drinking Water Samples Using the Chernoff's Faces Visualization Approach. Polish J. of Environ. Stud. vol. 15, no. 5 (2), pages 691–697.

Babulak E. (2006). Quality of Service Provision Assessment in the Healthcare Information and Telecommunications Infrastructures. Int J Med Inform 75, pages 246–252.

Ballard, C., Rollins, J., Ramos, J., Perkins, A., Hale, R., Dorneich, A., Milner, E.C., Chodagam, J. (2007). Dynamic Warehousing: Data Mining Made Easy, IBM.

Barnett, V. Lewis, T. (1994). Outliers in Statistical Data. John Wiley & Sons, 3rd edition.

Batagelj, V., Mrvar, A. (2006). Program for Analysis and Visualization of Large Networks, Reference Manual. University of Ljubljana, Ljubljana.

Ben-Dor, A., Yakhini, Z. (1999). Clustering Gene Expression Patterns. In Proceedings of the 3rd Annual International Conference on Computational Molecular Biology (RECOMB 99), Lyon, France, pages 11–14.

Berkhin, P. (2006). Grouping Multidimensional Data: A Survey of Clustering Data Mining Techniques, Springer Berlin Heidelberg.

Bohanec, M. (2006). Odločanje in modeli. Društvo matematikov, fizikov in astronomov, Ljubljana.

Bohanec, M. (2011). Program for Multi-Attribute Decision Making, User's Manual, Version 3.03. IJS Report DP-10707, Jožef Stefan Institute, Ljubljana.

Breiman, L., Friedman, JH., Olshen, RA., Stone, C.J. (1984). Classification and Regression Trees. Wadsworth, Belmont.

Breunig, M., Kriegel, H., Ng, R., Sander, J. (2000). Lof: Identifying Density-Based Local Outliers. In Proc. of SIGMOD'2000, pages 93–104.

Brown, D. (2007). Scientific Communication and the Dematerialization of Scholarship. http://www.csa.com/discoveryguides/scholarship/gloss_f.php.

Cadez, I., Smyth, P., Mannila. H. (2001). Probabilistic Modeling of Transactional Data with Applications to Profiling, Visualization, and Prediction. In Proceedings of the 7th ACM SIGKDD, pages 37–46, San Francisco, CA, USA.

Cleveland, W.S. (1993). Visualizing Data. Hobart Press, New Jersey.

Chandola, V., Banerjee, A., Kumar, V. (2009). Anomaly Detection: A Survey. ACM Computing Surveys (CSUR), no. 3.

Coiera, E. (2009). Research Paper: Building a National Health IT System from the Middle Out. Journal of the American Medical Informatics Association, pages 271–273.

Dhillon, I. S., Fan, J. Guan. Y. (2001). Efficient Clustering of Very Large Document Collections. Data Mining for Scientific and Engineering Applications, Kluwer Academic Publishers, pages 357–381.

Dietterich, T.G. (1997). Machine Learning: Four Current Directions. AI Mag 18(4), pages 97–136.

Dranove, D., Kessler, D., McClellan, M., Satterthwaite, M. (2002). Is More Information Better: The Effects of 'Report Cards' on Health Care Providers. Working Paper w8697, National Bureau of Economic Research, Cambridge.

Dresner, H. (2003). Business Activity Monitoring: BAM Architecture. Gartner Symposium ITXPO, Cannes, France.

Džeroski. S., Goethals, B., Panov, P. (eds.) (2010). Inductive Databases and Constraint-Based Data Mining. Springer, New York.

Ester, M., Frommelt, A., Kriegel, H., Sander, J. (2000). Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support. Data Mining and Knowledge Discovery, pages 193–216.

Fix, E., Hodges, J.L. Jr. (1951). Discriminatory Analysis, Nonparametric Discrimination. USAF School of Aviation Medicine, Randolph Field, Tex., Project 21-49-004, Rept. 4.

Foss, A., Wang, W., Zaane, O. (2001). A Non-parametric Approach to Web Log Analysis. In 1st SIAM ICDM, Workshop on Web Mining, Chicago, IL, USA, pages 41–50.

Freund, Y., Schapire, R.E. (1997). A Decision-theoretic Generalization of On-line Learning and an Application to Boosting. J Comput Syst Sci 55(1), pages 119–139.

Friendly, M. (2006). A Brief History of Data Visualization. Handbook of Computational Statistics: Data Visualization, Springer-Verlag, Heidelberg.

Han, J., Kamber, M., Tung, K.H. (2001). Spatial Clustering Methods in Data Mining: A Survey. Geographic Data Mining and Knowledge Discovery. Taylor and Francis, pages 1–29.

Han, J., Kamber, M. (2006). Data Mining: Concept and Techniques. Morgan Kaufmann Publishers.

Hao, Q., Xue, Y., Shen, W., Jones, B., Zhu, J. (2010). A Decision Support System for Integrating Corrective Maintenance, Preventive Maintenance, and Condition-based Maintenance, Construction Research Congress 2010, Banff, Alberta.

Hart, M. (1999). What are Sustainable Indicators? Keywords: Indicators of Sustainability. http://www.subjectmatters.com/indicators.

Haux R. (2006). Health Information Systems – Past, Present, Future. Int J Med Inform 75, pages 268–281.

Hawkins, D. (1980). Identification of Outliers. Chapman and Hall.

HCP. (2007). Euro Health Consumer Index. Health Consumer Powerhouse. The EHCI report 2007.

Heer, J., Chi, J. (2001). Identification of Web User Traffic Composition Using Multimodal Clustering and Information Scent. In 1st SIAM ICDM, Workshop on Web Mining, Chicago, IL, USA, pages 51–58.

Hodge, V.J., Austin, J. (2004). A Survey of Outlier Detection Methodologies. Artificial Intelligence Review, 22 (2), pages. 85–126.

Hunt, E.B., Marin, J., Stone, P.J. (1966). Experiments in Induction. Academic Press, New York.

Inselberg, A. Dimsdale, B. (1990). Parallel coordinates: A tool for visualizing multi-dimensional geometry. In Proceedings of the 1st IEEE Conference on Visualization, San Francisco, pages 361–378.

Keim, D. Panse, C., Schneidewind, J., Sips, M., Hao, M. C., Dayal, U. (2003). Pushing the limit in Visual Data Exploration. Lecture Notes in Computer Science, pages 37–51.

Keim, D., Andrienko, G., Fekete, J., Görg, C., Kohlhammer J., Melançon, G. (2008). Visual Analytics: Definition, Process, and Challenges. Lecture Notes in Computer Science, pages 154–175.

Kralj Novak, P., Lavrač, N., Webb, G.I. (2009). Supervised Descriptive Rule Discovery: a Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining. J. mach. learn. res., vol. 10, pages 377–403.

Ladner, S., Arnone, R., Sandidge, J., DongShan, K., Casey, B., Hall, C. (2009). "Ocean weather" in the Gulf of Mexico: Exploiting Real-time Satellite Ecological Properties and Circulation Models for Coastal Ocean Monitoring, OCEANS 2009, Marine Technology for Our Future: Global and Local Challenges, Biloxi, MS, USA.

Langton, J.T., Prinz, A., A., Wittenberg, D., K., Hickey, T. J. (2007). Leveraging Layout with Dimensional Stacking and Pixelization to Facilitate Feature Discovery and Directed Queries. Pixelization Paradigm, Lecture Notes in Computer Science, pages 77–91.

Lavrač, N., Bohanec, M., Pur, A., Cestnik, B., Debeljak, M., Kobler, A. (2007). Data Mining and Visualization for Decision Support and Modeling of Public Health Care Resources. Journal of biomedical informatics, vol.40, no. 4, pages 438–447.

MediMap (2004). Knowledge Management in Medicine and Healthcare (Analiza podatkov za upravljanje znanja na področju zdravstva). Project financed by the Ministry of Health of the Republic of Slovenia.

MediNet (2005). Analysis of Factors in Setting up a Network of Health Care Personnel (Analiza dejavnikov za postavitev mreže zdravstvenih delavcev). Project financed by the Ministry of Health of the Republic of Slovenia.

MediNet+ (2006). Development of a Primary Health-care Network in Slovenia (Izdelava modela mreže zdravstvenih delavcev primarne ravni Slovenije). Project financed by the Ministry of Health of the Republic of Slovenia.

MediNet++ (2008). A Web Page for Presenting the Primary Health-care Network in Slovenia (Internetna predstavitev rezultatov analiz za spremljanje mreže zdravstvenih delavcev primarne ravni v Sloveniji). Ministry of Health of the Republic of Slovenia.

MoH. (2000). Slovenian National Program of Health – Health for All to 2004. Ministry of Health of the Republic of Slovenia. Official Gazette RS 49/2000.

Münz, G., Li, S., Carle, G. (2007). Traffic Anomaly Detection Using k-means Clustering. GI/ITG Workshop MMBnet.

Nathiya, G., Punitha, S.C., Punithavalli, M. (2010). An Analytical Study on Behavior of Clusters Using K Means, EM and K* Means Algorithm. (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, no. 3.

NHPC. (2001). National Health Performance Framework Report. Queensland Health.

Nelson, G.S. (2010). Business Intelligence 2.0: Are we there yet? SAS Global Form, Chapel Hill, North Carolina.

Niven, R. (2003). Balanced Scorecard for Government and Nonprofit Agencies. John Wiley and Sons Inc., Hoboken, NJ.

OECD. (2008). Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD, Paris.

Patcha, A., Park, J. (2007). An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends. Computer Networks.

Pechter, R. (2009). What's PMML and What's New in PMML 4.0, ACM SIGKDD Explorations Newsletter, Carlsbad, CA, USA.

Pokrajac, D., Lazarevic, A., Latecki, L. J. (2007). Incremental Local Outlier Detection for Data Streams. IEEE Symposium on Computational Intelligence and Data Mining (CIDM).

Power, J.D. (2002). Decision Support Systems: Concepts and Resources for Managers. Quorum Books Division Greenwood Publishing, Westport, CT.

Pur, A. (2001). Upravljanje z znanjem s pomočjo expertnih sistemov: primer uporabe pri prečevanju kaznivih dejanje, Znanstveno posvetovanje o razvoju organizacijskih ved, 2001, Portorož.

Pur, A., Bohanec, M. (2003). Knowledge Discovery from Data Bases - Possibility of Using Association Rules in Police. Varstvoslovje, FPVV, no. 1, pages 16–21, Ljubljana.

Pur, A., Bohanec, M., Cestnik, B., Lavrač, N., Debeljak, M., Kopač, T. (2005a). Data Mining for Decision Support: An Application in Public Health Care. Innovations in Applied Artificial Intelligence, Lecture Notes in Computer Science, vol. 3533, pages 459–469.

Pur, A., Belič, I. (2004). The Telephone Traffic Data Analysis: Dilemmas of Contemporary Criminal Justice. The Fifth Bienniale International Criminal Justice Conference, 23–25 september, 2004, Ljubljana.

Pur, A., Bohanec, M., Lavrač, N., Cestnik, B. (2007b). Data Presentation Methods for Monitoring a Public Health Care System. 11th Mediterranean Conference on Medical and Biological Enginering and Computing 2007, 26-30 June, 2007, Ljubljana, Slovenia, (IFMBE proceedings, vol. 16). New York: Springer: International Federation for Medical and Biological Engineering, pages 708–711.

Pur, A., Bohanec, M., Lavrač, N., Cestnik, B. (2010). Primary Health care Network Monitoring: A Hierarchical Resource Allocation Modeling Approach, The International Journal of Health Planning and Management.

Pur, A., Bohanec, M., Lavrač, N., Cestnik, B., Debeljak, M., Gradišek, A. (2007a). Monitoring Human Resources of a Public Health Care System through Intelligent Data Analysis and Visualizatin. Artificial Intelligence in Medicine, Lecture Notes in Computer Science, vol. 4594, pages 175–179.

Pur, A., Pribakovič, B. R., Lavrač, N., Bohanec, M., Cestnik, B., Urbančič, T., Albreht, T., Kopač, T., Lukšič, P. (2005b). Sodobne metode analiziranja in vrednotenja primarne in sekundarne ravni zdravstvenega varstva = New methods in analyzing and evaluating primary and secondary health care. Bilt.-ekon. organ. inform. zdrav., no. 21, pages 30–31.

Quinlan, J.R. (1979). Discovering Rules by Induction from Large Collections of Examples. In: Michie D (ed.), Expert Systems in the Micro Electronic Age. Edinburgh University Press, Edinburgh.

Quinlan, J.R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo.

Rahm, E., Hai Do, H.(2000). Data Cleaning: Problems and Current Approaches. Data Engineering, Special Issue on Data Cleaning, no. 4.

Ramaswamy, S., Rastogi, R., Shim, K. (2000). Efficient Algorithms for Mining Outliers from Large Data Sets. In Proc. of SIGMOD'2000, pages 427–438.

Saaty, T.L. (1994). Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process. RWS Publications, Pittsburgh, PA.

Shneiderman, B. (1992). Tree Visualization with Treemaps: A 2D Spacefilling Approach. ACM Transactions on Graphics, vol. 11, no.1, pages 92–99.

Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information visualizations. Department of Computer Science, Human-Computer Interaction Laboratory, and Institute for Systems Research University of Maryland, Maryland.

Srikant, R., Agrawal, R. (1996). Mining Quantitative Association Rules in Large Relation Tables. IBM Almaden Research Center, San Jose.

Steinbach, M., Karypis, G. Kumar, V. (2000). A Comparison of Document Clustering Techniques. In 6th ACM SIGKDD, World Text Mining Conference, Boston, MA, USA.

Tan, P.N, Steinbach, M., Kumar, V. (2006). Introduction to data mining. Pearson Addison-Wesley.

Turban, E., Sharda, R., Delen, D., King, D., Aronson, J.E. (2010). Business Intelligence: A Managerial Approach. Pearson Education Canada.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer, New York.

Wang, X., Dou, W., Ribarksy, W., Chang, R. (2009). Visualization as Integration of Heterogeneous Processes. Visualization Center, University of North Carolina at Charlotte, Charlotte, USA.

WHO. (2000). How Well Do Health System Perform? World Health Report 2000: Health Systems Improving Performance, http://www.who.int/whr/2000/en/whr00_en.pdf, accessed January 26, 2008.

Widrow, B., Rumelhart, D.E., Lehr, M.A. (1994). Neural Networks: Applications in Industry, Business and Science.

Williams, J.G. (1999). The Use of Clinical Information to Help Develop New Services in a District General Hospital. Int J Med Inform 56, pages 151–159.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z., Steinbach, M., Hand, D.J., Steinberg, D. (2007). Top 10 Algorithms in Data Mining, Springer, New York.

Xu, R., Wunsch, D. (2005). Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, Vol. 16, No 3.

Zhanpeng, J., Yuwen, S., Cheng, A.C. (2009). Predicting Cardiovascular Disease from Real-time Electrocardiographic Monitoring: An Adaptive Machine Learning Approach on a Cell Phone. Engineering in Medicine and Biology Society.

Zouba, N., Brémond, F., Thonnat, M., Anfosso, A., Pascual, E., Mallea, P., Mailland, V., Guerin, O. (2009). A Computer System to Monitor Older Adults at Home: Preliminary Results. Gerontechnology Journal, pages 129–139.

List of Figures

List of Tables

Appendix: Biography

Aleksander Pur was born in Celje, Slovenia, on April 23, 1960.

He completed a master's of science degree at the Faculty of Organizational Sciences, University of Maribor. Afterwards, he enrolled at the Ph.D. programme New Media and E-science at the Jožef Stefan International Postgraduate School. His research work is focused on monitoring and assessment systems, decision support systems, and data analysis techniques.

He has joined Ministry of the Interior since 1991. He is a Police Superintendent in Applications Development Division. He is working in the development of document management systems, international data exchange systems, border information systems, data analysis systems, risk management systems, and E-learning systems.